Latent gaze information in highly dynamic decision-tasks.

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Benedikt W. Hosp
aus Klettgau-Grießen

Tübingen
2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	04.02.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Enkelejda Kasneci
2. Berichterstatter:	Prof. Dr. Ansgar Thiel

Acknowledgements

I would like to thank various people who supported my constantly new, crude ideas about gaze behavior and its exploration. First and foremost, my parents, Franziska and Werner, to whom I owe not only the fact that I was allowed to follow this path, but also a loving upbringing and a lot of motivation (which have made me the person I am today). Also, my siblings, Josy and Philipp, who have always been there for me and to whom I have always looked up to.

I would like to thank Enkelejda Kasneci, Oliver Höner, Florian Schultz and Shahram Eivazi from University of Tübingen and Peter Haddawy and Myat Su Yin from Mahidol University. In addition to heated technical discussions we were having from time to time, they were always open to my ideas and always on hand, even when things got tight.

A big thank you also goes to my colleagues from the working group. Nora Castner, Wolfgang Fuhl, David Geißler, Thomas Kübler, Björn Severin, Efe Bozkir, and Yao Rong. Thank you very much for your support, your instructions, your help, and your humorous cooperation.

Last but certainly not least, I would like to thank Lioba, who has been with me through all the ups and downs.

Abstract

Digitization is penetrating more and more areas of life. Tasks are increasingly being completed digitally, and are therefore not only fulfilled faster, more efficiently, but also more purposefully and successfully. The rapid developments in the field of artificial intelligence in recent years have played a major role in this, as they brought up many helpful approaches to build on. At the same time, the eyes, their movements, and the meaning of these movements are being progressively researched. The combination of these developments has led to exciting approaches. In this dissertation I present some of these approaches which I worked on during my PhD.

First, I provide insight into the development of models that use artificial intelligence to connect eye movements with visual-expertise. This is demonstrated for two domains or rather groups of people: athletes in decision-making actions and surgeons in arthroscopic procedures. The resulting models can be considered as digital diagnostic models for automatic expertise recognition. Furthermore, I show approaches that investigate the transferability of eye movement patterns to different expertise domains and subsequently, important aspects of techniques for generalization. Finally, I address the temporal detection of confusion based on eye movement data. The results suggest the use of the resulting model as a clock signal for possible digital assistance options in the training of young professionals. An interesting aspect of my research is that I was able to draw on very valuable data from DFB youth elite athletes as well as on long-standing experts in arthroscopy. In particular, the work with the DFB data attracted the interest of radio and print media, namely DeutschlandFunk Nova and SWR DasDing. All resulting articles presented here have been published in internationally renowned journals or at conferences.

Zusammenfassung

Die Digitalisierung durchdringt immer mehr Lebensbereiche. Aufgaben werden zunehmend digital erledigt und damit schneller, effizienter, aber auch zielorientierter und erfolgreicher erfüllt. Die rasante Entwicklung im Bereich der künstlichen Intelligenz in den letzten Jahren hat dabei eine große Rolle gespielt, denn sie hat viele hilfreiche Ansätze hervorgebracht, auf die immer weiter aufgebaut werden kann. Gleichzeitig werden die Augen, ihre Bewegungen und die Bedeutung dieser Bewegungen immer weiter erforscht. Die Verknüpfung dieser Entwicklungen hat zu spannenden Ansätzen in der Wissenschaft geführt. In dieser Dissertation stelle ich einige der Ansätze vor, an denen ich während meiner Promotion gearbeitet habe.

Zunächst gebe ich einen Einblick in die Entwicklung von Modellen, die mit Hilfe künstlicher Intelligenz Verbindungen zwischen Augenbewegungsdaten und visueller Expertise herstellen. Dies wird anhand zwei verschiedener Bereiche, genauer gesagt zwei verschiedener Personengruppen, demonstriert: Sportler bei Entscheidungsfindungen und Chirurgen bei arthroskopischen Eingriffen. Die daraus resultierenden Modelle können als digitale Diagnosemodelle für die automatische Erkennung von visueller Expertise betrachtet werden. Darüber hinaus stelle ich Ansätze vor, die die Übertragbarkeit von Augenbewegungsmustern auf verschiedene Kompetenzbereiche untersuchen sowie wichtige Aspekte von Techniken zur Generalisierung. Schließlich befasse ich mich mit der zeitlichen Erkennung von Verwirrung auf der Grundlage von Augenbewegungsdaten. Die Ergebnisse legen eine Nutzung der Modelle als Zeitgeber für mögliche digitale Assistenzoptionen in der Ausbildung von Berufsanfängern nahe. Eine Besonderheit meiner Untersuchungen besteht darin, dass ich auf sehr wervolle Daten von DFB-Jugendkaderathleten sowie von langjährigen Experten in der Arthroskopie zurückgreifen konnte. Insbesondere die Arbeit mit den DFB-Daten stieß auf das Interesse von Radio- und Printmedien, genauer, DeutschlandFunk Nova und SWR DasDing. Alle hier vorgestellten Beiträge wurden in international renommierten Fachzeitschriften oder auf Konferenzen veröffentlicht.

Contents

1. List of Publications 1
   1.1. Scientific Contribution . . . 3

2. Introduction 5
   2.1. Expertise Research . . . 9
   2.2. States of Confusion . . . 12
   2.3. Data-Driven Analysis . . . 13

3. Major Contributions 17
   3.1. Gaze Expertise Linkage . . . 19
      3.1.1. Soccer Goalkeeper Expertise Identification . . . 19
      3.1.2. Differentiating Surgeons' Expertise . . . 25
      3.1.3. Conclusion . . . 29
   3.2. Cross-Domain Generalization . . . 31
      3.2.1. Cross-Domain Expertise-Related Gaze Features . . . 31
      3.2.2. A Deep-Learning Approach for Feature Selection . . . 35
      3.2.3. Conclusion . . . 36
   3.3. Gaze-Based Assistance Timing . . . 38
      3.3.1. States of Confusion during Arthroscopic Surgery . . . 38
      3.3.2. Conclusion . . . 40

4. Discussions & Outlook 41
   4.1. Gaze-Based Expertise . . . 42
      4.1.1. Soccer . . . 42
      4.1.2. Orthopedics . . . 44
      4.1.3. Outlook . . . 45
   4.2. Cross-Domain Generalization . . . 46
      4.2.1. Expertise-Related Features . . . 46
      4.2.2. Deep Learning Feature Selection . . . 48
      4.2.3. Outlook . . . 49
   4.3. Gaze-Based Assistance Timing . . . 50
      4.3.1. States of confusion during arthroscopic surgery . . . 51
      4.3.2. Outlook . . . 52

5. Ethical Considerations 53

A. Gaze Expertise Linkage 55
A.1. Soccer Goalkeeper Expertise Detection Based on Eye Movements . 56
A.1.1. Introduction 56
A.1.2. Methods 60
A.1.3. Results 72
A.1.4. Discussion 79
A.1.5. Conclusion & Implications 81
A.2. Differentiating Surgeons' Expertise Solely by Eye Movement Features 84
A.2.1. Introduction 84
A.2.2. Related Work 85
A.2.3. Participants and Methods 86
A.2.4. Results 89
A.2.5. Discussion 92
A.3. A Study of Expert/Novice Perception in Arthroscopic Shoulder Surgery 93
A.3.1. Introduction 93
A.3.2. Related Work 94
A.3.3. Participants, Materials and Methods 96
A.3.4. Analysis and Discussion 99
A.3.5. Conclusion 104

B. Cross-Domain Generalization 107
B.1. Cross-Domain Shared Expertise-Related Gaze Features 108
B.1.1. Introduction 108
B.1.2. Methods 109
B.1.3. Results 112
B.1.4. Discussion 114
B.2. Expertise Classification of Soccer Goalkeepers in Highly-Dynamic Decision-Tasks: A Deep-Learning Approach for Temporal and Spatial Feature Recognition of Fixation Image Patch Sequences 116
B.2.1. Introduction 116
B.2.2. Methods 120
B.2.3. Results 128
B.2.4. Discussion 130

C. Gaze-Based Support Timing 135
C.1. States of Confusion: Eye and Head Tracking Reveal Surgeons' Confusion During Arthroscopic Surgery 136
C.1.1. Introduction 136
C.1.2. Methods 138
C.1.3. Results 140
C.1.4. Discussion 141

References 143

1. List of Publications

Published Articles

1. M. S. Yin, P. Haddawy, **B. W. Hosp**, P. Sa-ngasoongsong, T. Tanprathumwong, M. Sayo,and A. Supratak. "A Study of Expert/Novice Perception in Arthroscopic Shoulder Surgery." In Proceedings of the 4th International Conference on Medical and Health Informatics (pp. 71-77). August 2020. https://doi.org/10.1145/3418094.3418135

2. **B. W. Hosp**, F. Schultz, O. Höner, and E. Kasneci. "Soccer Goalkeeper Expertise Identification Based on Eye Movements." PloS one, 16(5), e0251070. 2021. https://doi.org/10.1371/journal.pone.0251070

3. **B. W. Hosp**, F. Schultz, E. Kasneci, and O. Höner. "Expertise classification of soccer goalkeepers in highly dynamic decision tasks: A deep learning approach for temporal and spatial feature recognition of fixation image patch sequences," Frontiers in Sports and Active Living, vol. 3, p. 183, 2021. https://doi.org/10.3389/fspor.2021.692526

4. **B. W. Hosp**, M.S. Yin, P. Haddawy, R. Watcharopas, P. Sa-ngasoongsong, E. Kasneci. "States of Confusion: Eye and Head Tracking Reveal Surgeons' Confusion during Arthroscopic Surgery." In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA. https://doi.org/10.1145/3462244.3479953

5. **B. W. Hosp**, M. S. Yin, P. Haddawy, P. Sa-ngasoongsong, and E. Kasneci. „Differentiating Surgeons' Expertise Solely by Eye Movement Features". Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA. https://doi.org/10.1145/3461615.3485437

Submitted Articles

1. **B.W. Hosp**, F. Schultz, O. Höner, and E. Kasneci. "In the Search of A Superior Gaze Behavior: Cross-Domain Shared Expertise-Related Gaze Features." Submitted to: ACM Symposium on Eye Tracking Research & Applications (ETRA '22).

Additional Articles

1. **B. W. Hosp**, S. Eivazi, M. Maurer, W. Fuhl, D. Geisler, and E. Kasneci. „RemoteEye: An Open-Source High-Speed Remote Eye Tracker." Behavior research methods, 1-15. 2020. https://doi.org/10.3758/s13428-019-01305-2

2. W. Fuhl, S. Eivazi, **B. W. Hosp**, A. Eivazi, W. Rosenstiel. and E. Kasneci. "BORE: Boosted-Oriented Edge Optimization for Robust, Real Time Remote Pupil Center Detection." In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (pp. 1-5). June 2018. https://doi.org/10.1145/3204493.3204558

3. W. Fuhl, E. Bozkir, **B. W. Hosp**, N. Castner, D. Geisler, T. C. Santini, and E. Kasneci. "Encodji: Encoding Gaze Data into Emoji Space for an Amusing Scanpath Classification Approach." In Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (pp. 1-4). June 2019. https://doi.org/10.1145/3314111.3323074

1.1. Scientific Contribution

This work is divided into five chapters that focus on different aspects of latent gaze information in highly dynamic decision-tasks. The first chapter lists the publications that are related to this work and contains an overview of their scientific contribution. The second chapter introduces the necessary basics by starting with an introduction into the fundamentals and then proceeding to expertise and confusion research, which are two important perceptual-cognitive processes for this work. Chapter three sets the focus on the main contributions, which are the objective, robust, and reproducible linkage between gaze and expertise, which all three are exemplary shown on two data sets from different domains. Another contribution shows an approach to infer expertise-related features that are shared between different domains as a step towards general gaze expertise definition. One further contribution is an automation step to remove arbitrariness and manual feature selection in the process of building a model for expertise detection. The last contribution is an online system to detect states of confusion that can be used to temporarily schedule assistance options. The results of the mentioned contributions are discussed in terms of applicability and transferability in chapter four. Chapter five focuses on ethical considerations regarding behavioral data and machine learning. This work is based on the papers from the upper publication list. The original publications are listed in the appendix.

2. Introduction

While human beings might never be able to read another person's mind, science is already able to deduce certain information about cognitive processes based on user monitoring and data-driven methods. Eye tracking is one of the most promising emerging technologies that provides these insights, as the movements of our eyes reveal a lot of information about our cognitive states that are not obviously visible, but more subtle. Eye tracking as a method has already been used in ancient times, but the technology of video-occulography has its roots in the mid of the 20th century. It evolved from a mere laboratorial technique that included sophisticated ideas about optic systems to observe the movements of the eyes (i.e. Yarbus experiment 1967, [1]), to a wide field of ubiquitous and precise devices from handful vendors [2], [3].

Lately, this technology has used video cameras to record the eyes and their movements to provide insights into perceptional processes. To date, several cognitive effects were researched by investigating movements of the eyes. Based on the cognitive load theory [4], i.e. scientists can reason the mental effort of a person during a task or understand the attention and mental effort while driving by examining the changes in pupil sizes [5]–[10]. Other studies, for example, deal with the influence of anxiety on cognitive load [11] or how online learning affects the mental load of students [12]. However, eye-tracking technology can be used to detect even more complex perceptual-cognitive processes. The detection of some of these processes can be advantageous for the optimized development of human behavior. Especially in fields related to subconscious behavior and high dynamics, behavior can be hard to describe. There are several aspects to be mentioned, but the two most interesting ones are expertise and confusional states. There is great interest in the recognition of these processes via visual perception [13]–[30], as the application of eye tracking as a research method can provide objective measures, and likewise, can improve objective diagnostics of expertise and confusion in several fields, like soccer or medicine, and thus, enabling an objective way to understand and analyze subtle behavior better.

Hardware

The most common types of eye-tracking devices are mobile (head-mounted) eye trackers and remote eye trackers. Mobile eye trackers are worn like glasses as

where remote eye trackers are placed in front of or under a computer display facing the subject in front of the screen. Next to commercial systems, there are a plethora of open-source eye trackers available. Multiple devices with different sampling rates as well as different accuracy or precision values (important attributes that describe the performance of an eye tracker) are available. One of my first contributions to the scientific community was the development of "RemoteEye" [31]. With over 500 Hertz and an accuracy of $< 1°$, I developed a low-cost, high-speed, and open-source remote eye tracker for research. The eye tracker is easy to handle and uses a feature-based approach. The most important aspect of this eye tracker is, that it can be built by anyone. It's parts are either off-the-shelf LEDs and cameras, 3D printed boxes, or simple aluminum bars. Besides creating a basic understanding of eye-tracking methods and algorithms for my later research, it was important to me to provide open and low-threshold access to this technology. In fact, open-source is very important for the research community, because eye tracking is still a niche market. While Tobii [32] started to produce game ready eye trackers, that can be used in several computer games and Microsoft included an API for eye-tracking devices in Windows OS, the usability for research is still limited. Current state-of-the-art commercial eye-tracking devices are very expensive or restrictive regarding data accessibility and therefore hinder a great number of researchers to use eye tracking as a method in their research. With "RemoteEye" [31], I developed a system that can easily be built and used for research even by small labs with a lower budget.

Gaze Signal

One of the aims of eye tracking is to tell where the subject is looking at in spatial and temporal dimensions. For this purpose, one popular approach is to capture the image of the eyes and find the center of the pupil and/or glints (also known as 1^{st} Purkinje images, [33]). Usually, glints are artificially created reflections on the cornea which come from infrared light-emitting diodes (LED) of the eye tracker. Infrared light is not visible to the human eye, so it does not blind or distract the subject. With the location of the pupil center and/or the glints, one can calculate the optical axis of the eye. To know the point of regard (POR) of the eyes one then needs to calculate either the visual axis of the eye or at least the relationship between POR and optical axis. This step is important, as the offset between optical and visual axis can deviate approximately 4-5° horizontally and 1.5° vertically [34], and differs from subject to subject. Thus, this is done by a person-specific calibration routine. While certain points in space or screen are shown to the subject, the locations of the pupil and/or glint centers are recorded. Subsequently, there are different options. One is modeling the offset between the optical and visual axis of the eye by computing a 3D model (illustrated in Fig. 2.1) of the eye.

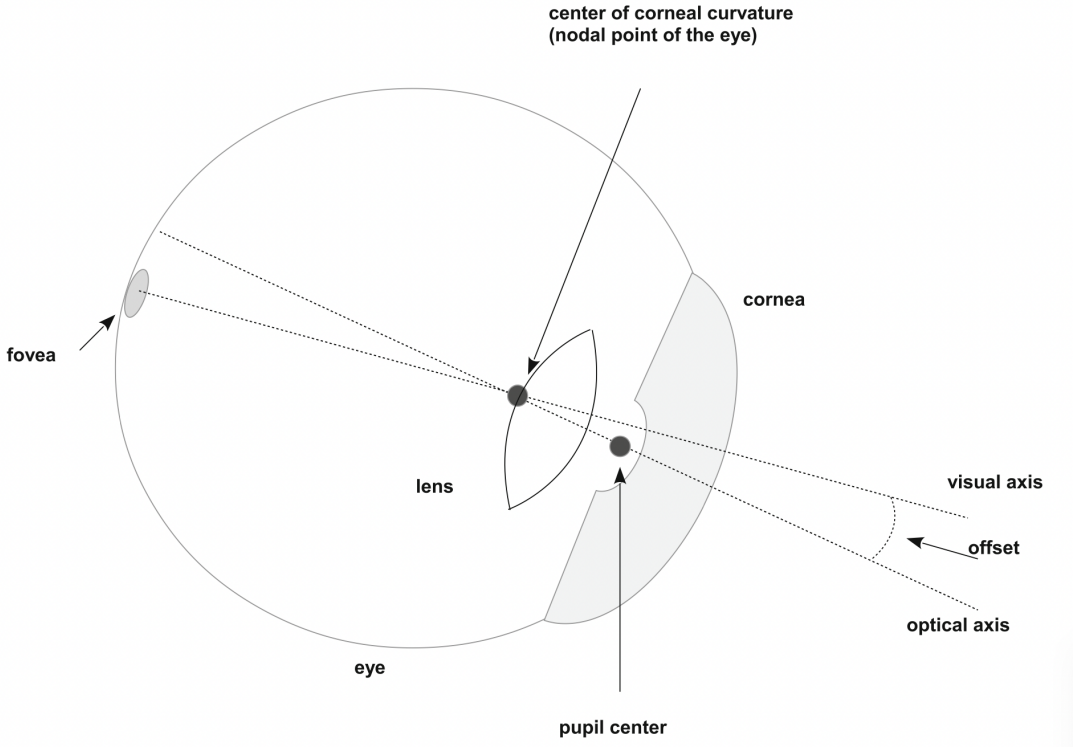

Figure 2.1.: Illustration of an eye showing important variables for gaze estimation. Source: Guestrin, Elias [35]

Another one is to use some combinations of pupil centers and/or glint features and the POG to estimate the relationship of the two axis, represented by an equation of a higher-order polynomial. This allows interpolating the relationship between POG and pupil/glint from any other location, usually with differing accuracy. It is the same technique for head-mounted and remote eye trackers, except that remote eye trackers need to find the face and eyes in the captured image first and are further away from the eye. Thus, they have fewer pixels to cover each eye but are usually connected to a powerful desktop computer, allowing higher processing speed. In the latest years, the common method of feature-based eye trackers (that are based on features like pupil center or glints), has slowly been replaced by uprising appearance-based methods. Appearance-based methods mostly take advantage of machine/ deep learning by taking images of the whole eye, finding and collecting features of the eye by themselves and relating them to certain points that are shown in space or screen during calibration. These techniques need a lot of computational power, but their accuracy is constantly improving and will run down the performance of feature-based methods anytime soon.

Eye Movements

On average, humans make 3-5 eye movements per second [33] which are needed to perceive visual information from our environment. Which eye movement events can be calculated resiliently, mainly depends on the speed of acquisition of the eye-tracking signal (in Hertz) [36]. In the center of our visual field, humans see the best when the object is imaged directly on the fovea which is an approximately 2° big area, slightly displaced at the back of the eye on the retina (as mentioned before, this is on the visual axis which can deviate up to 5° from the optical axis). Looking at the retina, one can see that the fovea is the point of highest acuity. Points that lie further away from the fovea, have less receptors. Therefore, the further away from the fovea, the worse vision becomes.

Fixations and saccades are seen as basic eye events. A fixation is theoretically a temporally and spatially limited accumulation of gaze points, which is calculated by an algorithm. It is assumed that visual information acquisition occurs during fixations. In detail, however, there are different methods with which a fixation can be computed. For low-speed eye-tracking devices (∼50Hz) threshold-based algorithms are mostly used (for overview see [37]). With higher speed, velocity-based algorithms are more often used [33] as they earlier detect saccade launches. However, there are other approaches, too, e.g. using bayesian statistics [38], [39] or even machine learning techniques [40].

Saccades are the jumps between fixations. That is, saccades are made when attention is drawn to another object. Thus, the attention-generating object is again on our fovea (area on the retina with the highest acuity). During saccades, humans are blind. However, the brain interpolates the images so that we are not aware of it. Saccades can be triggered consciously or unintentionally. However, for healthy people, both eyes usually move simultaneously and in the same direction. In addition to saccades, there are also micro-saccades, whose significance for perception has not been conclusively clarified. Eye-tracking software often calculates only fixations and considers all points between two fixations as part of a saccade. Besides micro-saccades, tremor and drift are considered to be part of a fixation (so-called fixational eye movements). The prevailing opinion about their usefulness is that these minimal movements help the eye to stay on target and prevent the trigger from disappearing by constantly refreshing the potential on the cells of the retina. Other eye movements are named but difficult to calculate and detect. Vergence stands for the adjustment of focus on objects in different depths. Here the eyes move in opposite directions. Much more interesting movements are smooth pursuits. These cannot be triggered consciously and manifest themselves as a fixation on a moving object. However, if the movement of the object becomes too fast, smooth pursuits are no longer applied but instead small sequential saccades. The last known eye movement is the vestibular-ocular reflex. Here the head and eyes move in opposite directions. This is comparable to a fixation of the eyes while the

head is turning.

Based on fixation and saccade calculations, second-order features can be derived. In this work, properties derived from and describing primary gaze events, like fixations, saccades, and smooth pursuits, are called second-order features. These derived properties are particularly important in expertise research [41], as simple comparisons between accumulated fixations or saccades offer little information [42]. Much more meaningful are data on velocity, acceleration, deceleration, frequency, duration, dispersion, and amplitude. With the help of these properties, a much more precise picture of eye movement characteristics can be drawn. Most high-speed eye-tracking device vendors provide the calculations of these features as exportable CSV-files. These high-speed features are especially necessary for highly dynamic decision-tasks. The subject does not only need to be precise and correct, but also fast. This means there is little time to perceive important clues. In highly-dynamic situations, it is therefore important to record even volatile movements of the eyes. These volatile movements are particularly present in expertise classification.

2.1. Expertise Research

Expertise is a qualitative measure that describes a person's ability to solve a certain task or area of work. Therefore, beginners have less expertise than intermediates or even experts. Becoming an expert usually takes years of training and practice of purposeful procedures that contribute to the solution of the task. As a qualitative measure, there is a difficulty in measuring expertise quantitatively. A certain amount of hours or years of experience in the field is often used as a measure. However, this excludes talented individuals. Since everyone develops at their own pace, people with a lot of talent and relatively little experience can have a higher level of expertise than people with a lot of experience but little talent. 'How to measure expertise objectively and robust' is one of the central research questions in visual-expertise research. The diagnosis and differentiation of expertise at different levels are particularly important for understanding the factors underlying expertise and its development. In order to study the development of a person's cognitive and motor skills, the 'expert performance approach' [43] is often used. It states that expertise is best revealed in a laboratory situation when the conditions are kept as realistic as possible. In sports, for example, physical education and training often cannot be made more intensive. Therefore, emphasis has been placed on improving cognitive factors in recent years, which have lately been recognized as advantageous but have not been trained adequately so far. However, an objective diagnosis is needed first and foremost.

In order to create a diagnostic model, usually known classifications of subjects are used first. This classification is done, for example through talent scouts or

competitions (both are common in soccer), but also through the status in an educational program or the number of hours one has already invested in the solution of a task (e.g. in medicine). In any case, a known classification model is needed, to develop a diagnostic model. As a second step, for each expertise class, data is collected from as many subjects as possible that represent this class. This is important to provide the machine learning algorithm with enough examples from which decision boundaries can be derived. On the one hand, this allows a broad picture of characteristics for the respective class, but also a robust boundary to separate classes from each other. On the other hand, a typical problem in expertise research is that the number of experts is small. This leads to the fact that studies dealing with expertise and its research have a small number of experts or subjects, which makes this data extremely valuable, but of limited use in terms of generalization. In addition to a small number of expert examples, an unnatural environment can lead to unwanted effects. For example, in sports psychology, more efficient gaze behavior during decision-making has already been linked to expertise. Unquestionably, experts show strengths in finding and interpreting relevant cues, but when it comes to the question of which gaze behavior features can describe the differences, the findings diverge. The results of some studies [44]–[49] suggest that expert behavior (more experienced, more talented, or more successful performance) is associated with an increased number of fixations. On the other hand, [45], [50]–[58] find no dependence of expertise on the frequency of fixations. Further, there are even studies that conclude that fewer fixations can be associated with expertise [57], [59]–[61]. A similar situation exists with the length of fixations and the reference to expertise. For example, [44], [45], [47]–[49] linked shorter fixations with expertise and [55], [57], [59], [60] linked longer fixations with expertise. Some studies even found no significant relationship at all [45], [46], [52], [53], [56], [58], [61].

The different results may lead to the conclusion that there is either no relationship between expertise and gaze behavior or that it was not found. However, this is a fallacy, because the studies mentioned did not pay much attention to the naturalness of the scene. The method of data collection plays an important role. For example, [62] conclude after a closer look at these studies that when the demands resemble a realistic situation in a soccer game, expertise may be associated with shorter and more frequent fixations. Similarly, [63] report that differences in expertise are much more emphasized when subjects are exposed to a highly realistic scene during data collection. Thus, so-called internal and external validity are both of high importance.

Regardless of these contradictions, expertise has been formed over years of experience and practice. On the one hand, it is assumed that experts develop their own optimal methods of perception by solving highly similar tasks for many years and optimize their perception in the process. On the other hand, it is assumed that there are certain commonalities in the experts gaze behavior. In addition to these

commonalities, the differences between levels of expertise are also of particular interest for research [42]–[49], [62]–[71]. This interest stems from the possibility of deriving insights into visual perception at different stages of development, but also from the possibility of using the knowledge of the commonalities and differences to define unique expertise levels. Diagnosing the correct expertise level (based on findings of visual perception research) can function as a basis for the development of possible assistance options. Both, in turn, can be used for a perceptual-cognitive tutoring or training system. Diagnostics are needed to determine the correct level of expertise of a subject at any given point in time. Likewise, to define the assistance options to determine the necessary progress for each of these levels of expertise to grant the subject to move to the next higher level of competence by learning new aspects of visual gaze behavior.

Different expertise classes show different similarities in gaze behavior so that a beginner needs another kind of assistance than an advanced user [41]. In recent years, the perception of experts has been investigated in various fields and tasks. Aspects of perception that allowed separation of expertise classes have often been found but were typically thought to be limited to a certain domain or task. While a look at the current research situation shows a mass of expertise research studies, only little inter-domain or inter-task work is done. Most work is somehow limited to a task or domain. For example, [72] shows that it is possible to transfer expertise from familiar tasks to semi-familiar tasks, but not to unfamiliar tasks. However, they took the same subjects for both tasks, which introduces a high risk of enabling recognition of subject-specific characteristics instead of expertise. Thus, while differences have often been found, only little is known about inter-task or at least inter-domain expertise that is transferable or generalizable. The problem of a missing generalizable feature set that works for more than one task or domain has yet not been addressed properly. So far, no dedicated set of traits was found that is better suited to recognize expertise than others. Therefore, previous study results could hardly or not at all be transferred to other studies and were always limited to one field, task, or data set [72]. However, since it is expected that experts in the same task exhibit certain commonalities regarding their gaze behavior, in a subsequent step, experts could also exhibit certain commonalities regardless the task or even domain. To investigate this hypothesis, studies are needed that evaluate the gaze behavior with the exact same methods, on different tasks or in different domains. The overall question is whether expertise-related features derived from the visual behavior are consistent across domains and whether experts of different domains share some visual strategy features. A superior set of perceptual properties would lead to a complete overturning of our understanding of perceptional expertise.

2.2. States of Confusion

Simultaneously, people's gaze behavior is sought to provide the deduction of even more latent characteristics. For example, it would be useful to recognize when a person needs assistance while accomplishing a task [73]–[83]. In addition to mental load or expertise, eye movements might serve as a proxy to detect perplexity or confusion during the completion of a task. First, one needs to define what is meant by "confusion". In fact, there are a lot of different definitions. Laymen use it less specifically, thus, different than health workers. For example, [84] say „symptoms and signs which indicate that the patient is unable to think with his customary clarity and coherence“ or „disorientation in time and place“ [85]. The use of the word is in fact so ambiguous, that in 1984, [86] even conducted a study to find out how medical doctors and nurses define confusion and which symptoms they consider for it. Depending on the field of the health worker the definitions were quite different. In this work, confusion is considered as a temporal state of disturbance that inhibits the continuation of the task, the definition of the Faber Medical Dictionary stating confusion as „a condition in which among other things there is a disturbance of consciousness" [87] is used.

Confusion can arise for a variety of reasons. For example, medical tasks are often lengthy and many steps have to be considered. However, especially in procedures such as arthroscopy, tissue can be very similar in many places, so navigating from the portal hole (entry point) to the surgical site can be extremely difficult and confusing. That means, confusion can also occur when the surgeon does not know where they are, how to proceed, or cannot recognize helpful visual cues such as specific bone formations. Even an incorrect rotation or orientation of the arthroscope camera can lead to confusion. In addition to the correct projection of the 2D output video from the arthroscope camera onto the 3D surgical area, the surgeon has to know where they are at all times in order to reach the correct surgical site in the body. During training operations on real human bodies, a supervisor always has to be present to show the trainee the correct way to proceed, in case of doubt. By automatically detecting confusion, adequate assistance can be provided digitally. In the best case, there is no need for a supervisor to stand next to the trainee, which not only saves money but also valuable time for the supervisor, who has typically little time for training, as they are usually also in charge of other operations, too. Enriching the arthroscopic camera image with helpful hints that enable the surgeon to progress without the need for a supervisor would be a simple, cost- and time-saving way of training young surgeons. For this process, however, the first step is to recognize when the surgeon needs assistance. In a second step, the necessary assistive steps have to be defined.

Although head and eye movement measurements are often used to detect cognitive processes, these measurements were never used in medicine to detect states of confusion or perplexity (as far as the author knows). Here, too, it can be assumed

that the combination of eye (and head) tracking and machine learning can offer a suitable methodology to deal with such a problem. With the knowledge of moments of confusion, person-specific assistance possibilities can be discussed, as next to expertise, there is high interest in understanding learning behavior [88]–[91]. As such, the absence of expertise during a task is of high interest, too. Though, learning behavior is typically highly subjective. This means that each subject has its own learning speed, its own ways of learning, and is at different levels of expertise. Here, too, many processes take place unconsciously so that verbal communication about the correct focus in the correct moment is particularly problematic and does not lead to the desired results. For example, subjects may be unsure how a task should be carried out further or are inactive during a task, which may indicate that the subject is not sure about their decision. Especially for laypeople and beginners, complex tasks can cause perplexity or confusion. Therefore, beginners have to learn how to find suitable visual landmarks and use them optimally. This can be taught verbally only to a certain extent so that even just showing an expert viewpoint would be much easier to achieve with far greater added value. Novices in training need a teacher, but at least in some situations, the teacher can be replaced by a training system. Due to the complex nature of some real-world visual search tasks, it can be really difficult for the expert to describe visual points. An ordinal example gaze overlay of an expert could be recorded and shown asynchronously to several novices, simultaneously.

Confusion can be expressed in many ways, which makes it difficult to define a single measure that could be used to identify it. In contrast to expertise, however, class differences are less taken into account here than person-specific reactions are evaluated. Since people learn differently and at different rates, and since patients' tissues can differ greatly, there is always either some adaptation to the current situation or a highly abstracted, generalized approach needed. However, in order to define a uniform measure for recognizing states of confusion, certain characteristics must have their validity and be recognized in as many cases as possible. A few approaches were discussed and applied in the research of confusional states. For example, longer fixations on task-irrelevant areas were associated with confusion [92]. Regarding input sensing, work has already been done with electroencephalogram (EEG), but mainly with so-called think-aloud protocols, [93]. Little research has been done on the detection of confusional states by using eye-tracking data, which might also depend on the high complexity of the gaze signal.

2.3. Data-Driven Analysis

Luckily, for a few years now, machine learning is a rising field, which helps to deal with highly complex data such as gaze data. Machine learning and deep learning techniques found their way into the analysis of eye-tracking studies. Like

a perfect fit, eye tracking creates a lot of data and machine learning usually works better with more data, which results in synergetic effects. Next to the analysis of eye-tracking data, these synergetic effects can be advantageous in the search for features that are thought to reflect expertise but are too complex to analyze with traditional methods. In fact, machine learning or, more precisely, deep learning simplifies the recognition of expertise, cognitive load, or, as already mentioned, confusion. Two areas of artificial intelligence are applied in this work: machine learning and deep learning. On the one hand, there are classical machine learning methods (shallow classifiers) like Support Vector Machines (SVM) [94], Random Forest [95], or Logistic Regression [96]. On the other hand, there are deep learning techniques like artificial neural networks, which are considered to be a type of machine learning, while both are part of artificial intelligence. Deep learning and machine learning are hard to differentiate, but one of the main differences between machine learning and deep learning is the ability to process unstructured data through artificial neural networks (ANNs). This is because deep learning through ANNs is able to convert unstructured information such as texts, images, sounds, and videos into numerical values. This extracted information is then used for pattern recognition or further learning. Among others, both are part of the field of artificial intelligence.

Nowadays, different forms of machine learning are used in different scientific fields. For example, there is a lot of research using traditional machine learning methods for expertise recognition, such as in dentistry [13], [29], microsurgery [22], [97]–[99] or in sports [15], [100]. Likewise, the number of applied deep learning methods is slowly increasing, too [30], [101], [102]. However, there are multiple challenges that need to be addressed when developing an approach to artificial intelligence. To find the right approach, the following points should be considered. Basically, the type of available data plays a role. If the data is available in unstructured form, deep learning methods can be used. Machine learning requires a certain structure to be available. Either the developer puts the data into a structured form or lets a neural network do this work. If the data is too complex to find patterns and relationships between them, deep learning is more likely to be applied. For example, images contain complex information. Here, information is organized in an unstructured way represented by vertical and horizontal pixels. For a classical machine learning approach, it is necessary that each sample is described by defined attributes. Classically, tables are created that show the expression of certain specified features for each sample. Thus, machine learning needs a lot of pre-processing to work on images, which means that machine learning needs more intervention of the developer, whereas deep learning has a certain autonomy. The approaches also differ in the amount of time they require. Machine learning methods can be set up and executed quite quickly, but their expressiveness can be limited. Deep learning methods require more time to set up, but can usually deliver better results with more time as more data becomes available.

If one wants to apply these methods in eye tracking, the eye-tracking signal can be used directly in order to generate a structured representation to use classical machine learning methods. The eye-tracking signal is stored in a file (or can be retrieved online) that shows a timestamp and the current gaze point at that time on the stimulus. Based on this gaze signal, higher layer features like fixations and saccades can be computed. Several derivations like the frequency of the fixations and saccades or the velocities can then be calculated, too. The computation of such features allows the structured description of individual samples, which can be the statistics of the features for instance over a trial or a stimulus. Such a structured representation could contain features as columns and the respective expression of the features of a sample as rows. In order to train the algorithm, each sample of the training set is assigned to a class, too. For such a structure, classical machine learning methods can be applied. Usually, the goal is to obtain a representation of the gaze behavior that describes the behavior as good as possible in a structured way. The machine learning algorithm tries to identify similarities between samples of the same class and differences between samples of different classes and thus to separate the classes, by defining a decision boundary between them. Based on this, a model is created that learns the separation of classes as robustly as possible that new unlabeled samples are correctly classified. However, gaze behavior analysis can be done in many ways. For example, if the gaze behavior is not only in the form of a gaze signal but also in the form of images (e.g. the sequence of a scan path represented by AOIs or image portions of the stimulus), deep learning approaches can be used, since images are unstructured data, too. Such approaches have found application in current research. For example, if one assumes that expertise results from the optimal perception of helpful visual cues [67], it may be more useful to analyze behavior based on the sequence of visual cues. In the past, next to traditional algorithms such as Needleman-Wunsch [103] or Smith-Waterman [104], partly borrowed from bio-informatics, several types of scan path comparison algorithms have been developed [13], [18], [20], [21], [23]–[26], [29], [30], [105]–[107]. However, deep learning methods show a particular strength here. Which features in the images or videos (which are ultimately only a sequence of images) are used for classification can be determined by applying different layers in a neural network. Likewise, different filters can be used to determine any latent features in the image sequences. Whether this is based on the saliency of parts of the image, particular edge detection, or even object detections, is part of investigations of current research and is therefore left to the developers. For example, CNNs have shown that the convolution operation, after which CNNs are named, can extract an extremely large amount of information from an image by interleaving several operations. In most cases, the first layers of a neural network are designed to recognize edges, corners, patterns, and objects. At the time of writing this thesis, research on the optimal use of CNNs and optimization by residuals, and 3DCNN, was in full swing.

3. Major Contributions

This chapter summarizes the main contributions during my PhD work. For each of the papers presented in the following, I will describe the motivation for the research question and give a summary of the main findings. The full text of the papers can be found in the appendix. Figure 3.1 provides an illustration of the several pieces of work and how they are connected.

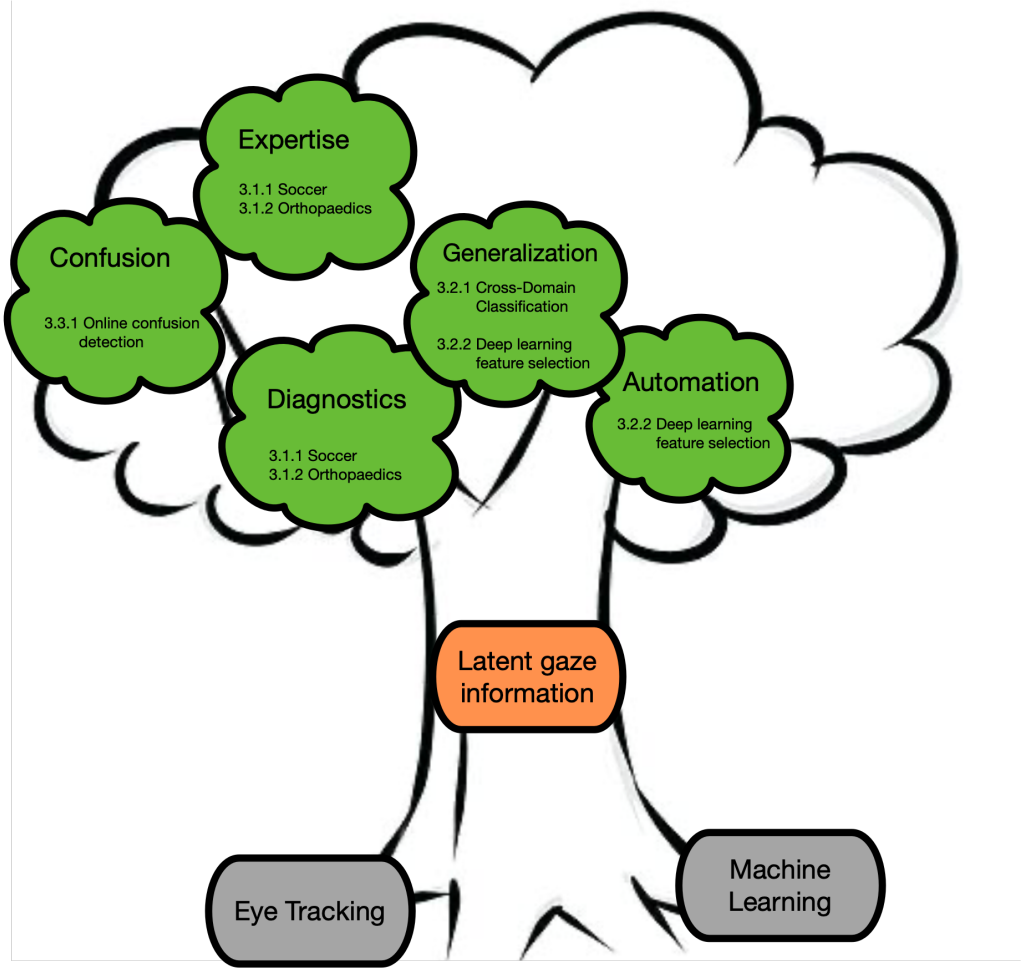

Figure 3.1.: Overview and interrelationships of the presented papers.

As foundational to this work, eye tracking and machine learning can be seen as the roots of a tree. Eye tracking allows vital access to conscious as well as subconscious gaze information. Machine learning opens up novel ways of dealing with its complexity in order to infer further, deeper knowledge from it. The quality of the gaze signal as well as the robustness and applicability of the machine learning methods are essential for research on latent gaze information. The more these two research fields grow (higher recognition rate, more robust results, new methods), the more possibilities for researching on latent gaze information are provided. As the foundation grows, it delivers more essential knowledge and procedures and thus allow to infer even more latent gaze features. Research on latent gaze information can be considered as the main trunk of the tree. Since it combines the powers of both root technologies, eye tracking and machine learning, it allows to build models to answer a pool of research questions. In turn, the answers to research questions can lead to new applications (new branches of the tree). With growing roots, the trunk of the tree can grow bigger and thus enable more new branches to grow. Each branch symbolizes applications one can build upon the research on latent gaze information. One of them is diagnostics. Diagnostics are essential for further work like training or tutoring systems. Aspects that are typically diagnosed are expertise (covered in section 3.1.1 and 3.1.2) and confusion (covered in section 3.3.1). By combining both aspects in a training system, not only the expertise level, and thus, the level of assistance needed (expertise detection), but also the correct timing for assistance can be found out (confusion detection). Another branch is leading to a unified design process which can be achieved by a certain degree of automation (covered in section 3.2.2). By focusing on the removal of manual selection, the arbitrariness is cut out of the process, which enables some kind of automation. For generalization, some kind of robust, objective, and reliable cross-domain diagnostics is needed, that is based on the same, unified and automated processes. This work includes first steps towards a general perspective on latent gaze information, their relationship and dependencies to diagnostics (covered in section 3.2.1).

3.1. Gaze Expertise Linkage

In certain research areas, gaze behavior has already been linked to expertise. However, this was mainly done manually, visually, or based on statistical methods [18], [20], [21], [23], [25], [26], [105], [106], [108]. Since a few years, machine learning techniques are commonly used to infer knowledge about scan paths [13], [22], [24], [29], [30]. In fact, machine learning methods can remedy this linkage by using supervised methods that lead to a uniform approach on the one hand, and to explainable results on the other. Building uniform methods to analyze gaze-based expertise is an important step for comparisons of expertise and its definition. This first contribution is an objective, reproducible machine learning approach, that results in a model with explainable features. Likewise, this model helps to understand the evolution of perceptional features in several stages of expertise. This approach is applied to two different groups of subjects - containing expert, intermediate, and novice subjects - from sports and medicine. The analysis of both data sets shows that it is possible to recognize expertise based on a few gaze features presented as a ternary classification problem, with high accuracy (78.2% and 76.46%, respectively). Further, the high influence of idiosyncrasy of human gaze behavior on classification is shown and, likewise, which features describe the differences in expertise the best.

3.1.1. Soccer Goalkeeper Expertise Identification based on Eye Movements

B. W. Hosp, F. Schultz, O. Höner, and E. Kasneci. "Soccer Goalkeeper Expertise Identification Based on Eye Movements." PloS one, 16(5), e0251070. 2021.

Motivation

While connections between expertise and gaze behavior have been made multiple times in different fields, little is done in sports, that 1) shows high accuracy on the classification of gaze behavior of three classes of expertise, 2) uses objective, reproducible, and state-of-the-art methods for classification, and 3) lead to explainable features. The following paper describes how to find an optimal set of features and feed it to a supervised machine learning algorithm to express the commonalities and differences of expertise-dependent gaze behavior in a robust, objective and reproducible way. The use of a classification model as an online diagnostic system

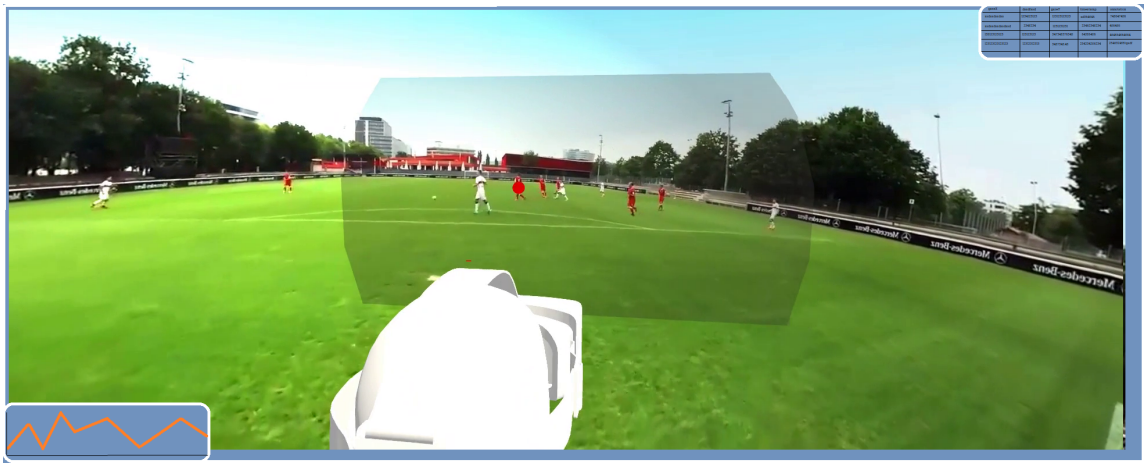

Figure 3.2.: Zoomed out example of subjects' perspective during data collection.

is one of the far-reaching aims of this work. Therefore, how these findings can be used in the future is discussed in chapter 4.1.

Methods

To infer the connection between gaze and expertise, a 360° camera was placed on the soccer field while soccer players physically replayed a defined common scene (Fig. 3.2 shows a zoomed-out perspective view of the subjects). This omnidirectional video footage was then presented to our subjects on virtual reality (VR) glasses. Each scene shows a build-up situation that ended by the return pass to the position of the subject. Afterward, the subjects had to tell how to continue the scene. In total,there are 33 subjects from three different expertise classes. While the subjects were watching the stimuli on VR glasses, their eye movements were captured with 250 Hz. In the first step, the main model with all eye-tracking features of the eye tracker available is constructed. In a second step, a subset of features was defined that increases the accuracy of a test set while reducing the number of features dramatically. Three different methods were investigated. All features that have the highest p-values, thus, significant differences between the three classes of expertise (significant features = SF), all features that show the highest frequency by ranking them with a maximum relevance minimum redundancy (MRMR) algorithm and chi-square-test (most frequent features = MFF) and for comparison a model with all features available from the eye tracker (all features = ALL).

For all of the approaches, one first needs all available features from the eye tracking device. All of them are taken into account to build a first model. The effective aim is to find an optimal yet small set of features that has a high impact

on the classification. A smaller feature set means shorter computing time and therefore better usability in an online diagnostic system. Starting with all features available, an SVM model is build to classify the data into three groups. The model reaches a certain accuracy Acc_m, which is the performance metric that is used to compare the impact of the features sets that were defined. Though the complete elimination of statistical errors is impossible, SF and MFF approaches need to be applied in a high number of runs, to lower the probability of statistical errors occurring.

Results

Feature set evaluation

One of the main findings is the comparison of the different subsets of features. The subset of features chosen by the MFF approach (MRMR and chi-square-test) shows better performance in prediction accuracy (78.2%) than a model with ALL(75.08%) or SF (73.95%). When looking at the 75th percentile, the differences are better noticeable (ALL: 80.989%, SF: 79.25%, MFF: 85.44%). A classification model is considered as well performing with an accuracy of over 70%. Thus, all three models can be considered as a classification model, but by looking at the recall of all three models, the MFF is the best performing, again (ALL: 71.87%, SF: 73.19%, MFF: 76.18%).

Idiosyncrasy

When assigning samples to the training and evaluation set, one has to consider an important point. Most eye movement features are idiosyncratic [33]. Actually, a large portion of eye movements have already been proven to be idiosyncratic, like fixation duration, blink duration and rate, pupil diameter, saccade acceleration and deceleration, and saccade amplitude. During the model training step, these findings could be approved. By randomly assigning all samples of all subjects to the training or evaluation set, samples of each subject end up being distributed on both data sets (Fig. 3.3). This leads to an unexpected learning behavior of the model, as the model rather matches the origin of the sample to a specific subject. Thus, it is not classifying a sample's class directly, but rather through its belonging to a certain subject. Such a model would estimate all samples of the evaluation set nearly perfectly, as the training set already contained highly similar samples of the same subjects. However, the model would fail to predict the belonging of new samples from new subjects correctly, as it has never seen data of this subject before. This behavior of idiosyncratic eye movement features reveals that the differences between subjects are much bigger than the differences within subjects. As such, a

classification model would learn a subject-specific, bio-metric relationship instead of a correct class representation.

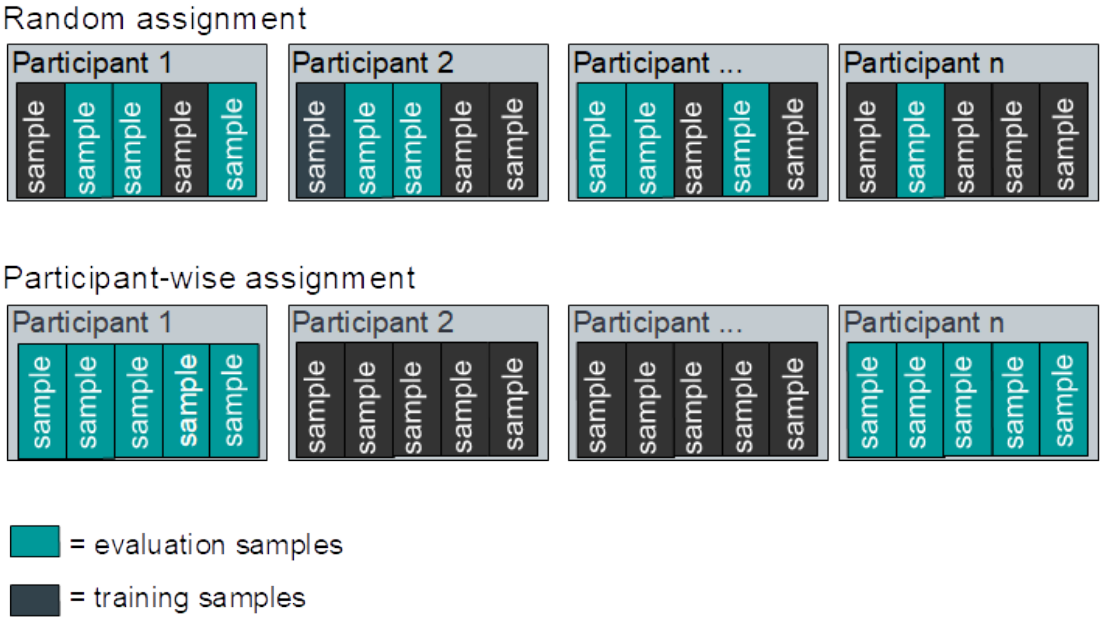

Figure 3.3.: Random and subject-wise (idiosyncratic proof) sample assignment.

Expert variation

A similar approach has been used to infer the classifiability of a certain group of subjects. Half of all experts were taken and switched with the same amount of intermediates by intentionally labeling them with the other class label. The accuracy is expected to drop under chance level, which would prove, that the differences between experts are smaller than the differences between experts and intermediates. The assumption was correct, as the model could not anymore differentiate between true experts and fake experts and intermediates vice versa. Defining a robust decision boundary was not possible anymore, which allows the statement to be defined: Differences between experts are smaller than differences between experts and intermediates.

Classification

In a further approach, the expert and intermediate samples are considered to build a first SVM model with all features, that is able to predict the affiliation of the samples. A classification accuracy of 88.1% was achieved. This means the model

is able to estimate the affiliation of a sample of an expert or an intermediate correctly, with a probability of 88.1%. There were 31 samples out of 260 falsely classified and the performance on intermediate samples was better than on expert samples. With a low miss rate of 11.9% the model shows great results. In a second approach, a better-performing model that needs fewer features was investigated. As the MFF set showed superior performance, this set is used in a ternary classification. As stated earlier, the ternary classification with the MFF set peaked at 78.20%. Compared to the chance level of guessing, this model can be considered to be performing outstandingly.

Latent expertise features

Next to the applicability of the MFF set as a foundation for a high-performing ternary classification, the MFF model revealed a certain amount of latent gaze information that is reflected by characteristics of the features. Most of the found features are typically not used as expertise markers. This might come from their difficult interpretability, as there is no obvious and simple characteristic behind these features. One first difference is found in the saccadic movements. Experts, as well as novices, tend to have a more homogeneous saccade behavior as the standard deviation of their saccade lengths is much smaller. However, novices have similarly long saccades as where experts have similarly short saccades. Apart from that, this allows proving the statement of [67], that experts have fewer but longer fixations. Their behavior is usually based on longer fixations to avoid saccadic suppression, as there is no information intake during a saccade. In this work, differences in fixation length between expertise groups were not directly found. This might be based on the split between short fixations and smooth pursuits or from the age difference between the single expertise groups. Conversely, further differences are found in the maximum deceleration of the saccades. In line with [109] deceleration behavior was found to be an adequate marker for expertise detection, too. There is a continuous increase in the maximum deceleration speed of the saccades. Novices are much slower than intermediates and experts.

Another quite interesting observation during data collection was the gaze behavior when the ball is passed around in the stimulus. Experts tend to only look at the ball shortly before and after a teammate is in possession of the ball. Novices tend to follow the track of the ball a much longer time. This is an important behavior as there are optimal times when the player can seek an overview over the scene to be able to react appropriately when getting into possession of the ball. These times are when the player is not in play, when the ball has been passed and cannot change its track, and when the line of sight to the ball is blocked (the subject is not playable). The values of the smooth pursuit dispersion vigorously prove such behavior. Experts have a small window between minimal and maximal smooth pursuit dispersion. Their maximum is less than half as long as the novices' and

their minimal value is still $1/3$ shorter than for novices. Intermediates are placed between experts and novices. Thus, there is a continuous decrease visible. Likewise, the average smooth pursuit, as well as the maximum and standard deviation of the smooth pursuit dispersion correlate negatively with the classes. The classes differ significantly, which is also reflected in the average, minimum, and maximum smooth pursuits, with a p-value of $p < 1 \times 10^{-12}$. Novices show much longer smooth pursuits than intermediates and experts. Likewise, the shortest smooth pursuits of the novices are longer than the intermediates' and experts'. The same patterns can be observed in the maximum values of the smooth pursuits, as novices have a higher maximum than the intermediates and the experts. The standard deviation of the lengths of the smooth pursuits shows a highly similar pattern but is statistically not significant. Novices' smooth pursuit scatter much more than intermediates or experts.

3.1.2. Differentiating Surgeons' Expertise Solely by Eye Movement Features

B. W. Hosp, M. S. Yin, P. Haddawy, P. Sa-ngasoongsong, and E. Kasneci. „Differentiating Surgeons' Expertise Solely by Eye Movement Features". Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA.

and

M. S. Yin, P. Haddawy, **B. W. Hosp**, P. Sa-ngasoongsong, T. Tanprathumwong, M. Sayo,and A. Supratak. "A Study of Expert/Novice Perception in Arthroscopic Shoulder Surgery." In Proceedings of the 4th International Conference on Medical and Health Informatics (pp. 71-77). August 2020.

Motivation

To infer how surgeons perceive their environment and especially how they navigate through tissue during arthroscopic surgeries, we investigate the distribution, analysis, and comparison of gaze behavior with a mobile eye tracker during live surgery. We want to know if, and when which, features in their gaze behavior can be found that differentiate different classes of expertise. Both papers are based on the same study. The work of [110] can be seen as a pilot study, where we focus on typical measures like area of focus distribution, cognitive load, and the classifiability of expertise with common gaze features. In [41] we focus more on the classification as we dig deeper into the classification of three groups of expertise. To further understand perceptual differences we focus on the developmental steps between the expertise groups, too. In both studies, we use the same recordings of the subjects as the data source. These findings are meant to be a base diagnostic system for a future training system that helps to understand the gaze behavior of surgeons in general and improve the education of young surgeons.

Methods

The gaze signal of 15 surgeons was recorded during a surgery on a soft-cadaver at the Ramathibodi Hospital in Bangkok, Thailand. There were n=5 experts with multiple years of experience, n=5 4th-year residents and n=5 3rd-year residents. All of them had to navigate from the portal on the shoulder to the operating site near the tendon of the shoulder, using an arthroscope (see Fig. 3.4 for an overview

Figure 3.4.: Surgeon during data collection looking at the output screen of the arthroscope.

of the scene). The surgeons had to tell when they were reaching one of 12 landmarks, which are placed on the way to the operating site. Starting from a rather general perspective, the visual attention of surgeons from different expertise levels was first focused on. ArUco markers are attached around the output screen of the arthroscope (4k, 52-inch, placed 4 feet away from the surgeon) to detect it and the circular video feed of the arthroscope therein, from the field camera video of the eye tracker (Tobii Glasses 2). The fixation patterns of experts and novices in the inner and outer circles of the arthroscope video are investigated, in order to state differences in the distribution of focus (Fig. 3.5). In the next step, the importance of confusion detection during navigation is emphasized. While there are plenty of visual clues within the joint to detect landmarks, novices often miss to diagnose the correct target landmark and thus, report times of confusion. Previous studies have already proven the connection between confusion or disorientation with change in pupil diameter and head movements. With the percentage of change in pupil diameter (PCPD) an objective measure of cognitive load is used. For example, [111] found higher cognitive load associated with higher PCPD values. Higher cognitive loads are suspected in times when a novice is confused, too, and considered the end of the last landmark to the beginning of the search for the next landmark as the baseline with low cognitive load, as the surgeon is not navigating during this

time phase. The PCPD is computed by subtracting the average diameter of the confused time from the baseline diameter and dividing it by the baseline diameter. In a further investigation, whether there are differences in expertise visible in common eye movement features will be inferred. For a deeper analysis, the importance values of each feature is used to create a ranking (MRMR and chi-square-test), which tells us about their impact on the accuracy of new data. To infer more information about the differences of expertise reflected by their gaze signal which are considered to help to improve the understanding of the differences, the subset of features that ranked the highest are looked at and their characteristics are focused on as an explanation for evolutionary steps between classes of expertise.

Results

Area of focus

80% of the navigation process from one landmark to another is considered to belong to a general area search. The remaining 20% are considered to belong to a zeroing in or fine adjustment process. During general area search, experts, as well as novices, tended to focus on the outer circle by a ratio of 2:1. The differences can only be seen in the fine adjustment phase of navigation, where experts shifted their focus to the inner circle by a ratio of 2:1, but novices still focalize on the outer circle area with a roughly similar ratio as before. This finding is interpreted as an indication that experts adjust their attention according to the portion of the navigation task, as they know how close they are to the desired landmark, while novices might not be able to tell precisely.

Figure 3.5.: Visualization of inner and outer circle on the arthroscope's output screen.

Cognitive load

Confused novices took up to two times longer than other novices to diagnose landmarks. In terms of time taken to accomplish the surgery, experts needed the least amount of time with the least standard deviation in task times. The pupil diameter changed during times of confusion by 1.02% (left eye) and 1.12 % (right eye) to the baseline. A much more extensive change can be seen in the gyroscope and accelerometer values.

Classification

For a classifiability test typical metrics are used that were often used in medical studies like fixation rate (Hz), saccade rate (Hz), fixation duration (ms), saccade duration (ms), and average time to first fixation (ms) [112]–[116]. In total, there are 12 metrics used (inclusive AOI intersections) that are extracted from the gaze data. Out of 15 subjects, two had an erroneous gaze signal from time to time, which would contaminate their overall statistics. Especially absolute statistics like fixation and saccade rate are not any more representative features of the samples of such subjects. Therefore, these samples were not used.

Depending on their gain ratio, the features with the five highest values for our classification model are picked. Our logistic regression model shows a great accuracy of 84% in classifying experts and novices. Only one expert has been misclassified as a novice, as their fixation rate is highly similar to that of novices and one novice has been misclassified due to similar characteristics to the experts in time to the first fixation.

Since the common eye-tracking features from the binary classification did not work for a ternary classification, alternative approaches for feature selection were searched for. For the further investigations on the classification of three groups of expertise, the importance of each of the features is calculated (with MRMR/chi-square-test) during classification and thereby get their impact on the accuracy. With a small amount of four features, one can classify the three classes of expertise with 76.46% accuracy. Why a reduction of features is meaningful, is extensively discussed in Section 4.1.1.

Latent expertise features

Finally for the sake of explaining differences between expertise groups the following features were found most important to differentiate the two groups of novices and experts.

- Average time to first fixation (ms)
- Fixation rate (Hz)

- Fixation rate AOI inner circle (Hz)
- Average fixation duration (ms)
- Average saccade duration (ms)

For the deeper analysis of the classification of three groups of expertise, the following features are found to be of interest:

- standard deviation of saccade peak velocity
- minimum saccade amplitude
- total saccade amplitude
- min gyroscope z

The binary model as well as the ternary model perform well and can therefore be well used to classify expertise. A deeper look at the characteristics of the four most frequent features from the ternary model shows that experts tend to have a more uniform distribution of peak velocities of the saccades. Thus, one interpretation of this result is that experts have a more structured speed behavior, which is more like a fixed scanning behavior. Higher values in the variance of the saccade peak velocity could signal a more chaotic gaze behavior, but it is hard to draw a conclusion. When looking at the minimal saccade length, a similar ratio is visible. Experts tend to have bigger minimal saccade lengths compared to intermediates and novices. In both features, the novices are between the experts and intermediates, which is atypical. Especially because the total amount of saccade amplitudes shows a reasonable evolution. Experts have a lower value than intermediates (who do more than twice the experts) and novices. Novices' scan path is more than five times higher than the scan path of experts and nearly double the scan path of intermediates. The last important feature is the gyroscope minimum measure on the z-axis. Z-axis can be seen as the movement from left to right of the head. Here, again, one sees an atypical behavior as intermediate subjects have the lowest minimum value followed by the experts and the novices.

3.1.3. Conclusion

To conclude the chapter about gaze to expertise linkage, the previous two papers are looked at individually and combined.

The work on soccer data set shows how important idiosyncrasy is, when distributing samples of subjects on training and testing data set. We could confirm that a mass of gaze behavior features underlies a certain idiosyncrasy, which leads

to overly positive results, when not taken into account properly. Further, the results could show, that the differences in gaze behavior between subjects from the same class are much smaller than the differences between subjects from different classes. In the work on surgeons, indicators of different search behaviors were found. While experts tend to have an optimized search behavior, novices seem to have more problems, thus, their behavior is more chaotic and less precise. This corresponds to the cognitive load values of novices and the completion times, which were two times higher than for experts.

Both data sets allow an objective, reproducible and robust classification of expertise, which can be seen as the basis for a diagnostic system. To robust such a system, more data of more subjects is needed. While the results of the work on soccer mainly highlighted the importance of different features describing the smooth pursuits, the results of the work on surgeons found a high influence on the length of the saccade amplitudes. However, the feature sets of both have one feature in common. Both pronounced the difference in the standard deviation of the saccade peak velocity. This feature can be understood as the variation of peak velocities of the saccades. A small value would show a homogeneous behavior, while a high value would show more chaotic behavior. In both data sets, one could see that experts have a much smaller standard deviation of saccade peak velocities. This means their behavior is less chaotic. This difference leads to the question of whether there are more commonalities between experts of different fields or if it was found by chance. This will be illuminated in the next chapter.

3.2. Cross-Domain Generalization

When talking about perception and perceptual expertise, one usually talks about it in certain limits like domain or task. So far, there has been no proof, that perceptual expertise is restricted to a domain. Thus, perceptual expertise can also be some kind of domain-independent talent, which is considered to be one aspect of a successful generalization. In the first work presented in this chapter, the question asked was whether a subset of features can be found, that - applied in different domains - elicits expertise domain independently. The same feature set has been used to infer expertise classes by training a model with one data set of one domain and testing the model with unknown data from the data sets of a second and third domain.

However, generalization has multiple aspects. Another investigated research question was whether one can remove arbitrariness out of the way of finding spatial and temporal features by focusing on simple features like fixation image patches. This would allow a certain degree of automation from which generalization could benefit, too. It should be possible to apply our approach to any other domain where some kind of scan path can be created to infer expertise classes and differences.

3.2.1. Cross-Domain Expertise-Related Gaze Features

B.W. Hosp, F. Schultz, O. Höner, O. and E. Kasneci. "In the Search of A Superior Gaze Behavior: Cross-Domain Shared Expertise-Related Gaze Features." Submitted to: ACM Symposium on Eye Tracking Research & Applications (ETRA '22).

Motivation

So far, there is no proof that a superior set of shared features exists that explains expertise on more than one domain or task. The following work focuses on the research of commonalities between different domains. To prove that there is a superior gaze behavior that is valid in multiple domains, there needs to be more research on cross-domain expertise, but one step into this direction has been done in this work by using a uniform, objective, and robust way of the feature selection process and apply this approach to at least three data sets of different domains or tasks. Such a subset of features would allow general statements about perception to be made, independent of domain. This approach needs to be defined and applied to find cross-domain but class-related expertise differences, reflected by a set of features. The current scientific understanding of perceptual expertise is mainly domain- or task-related, but there is no proof that these limits exist. The contribution of this paper contains the investigations on the generalization of perceptual

expertise, where indicators are presented that cross-domain commonalities exist.

Methods

In the first step all the gaze data from three studies that are accessible are collected. Data set A contains the samples of 33 soccer goalkeepers from the study in Section 3.1.1. Data set B contains all the samples from 15 subjects from the study in Section 3.1.2. Data set C is coming from a more static task. In data set C data of 58 dentists was collected during an OPT analysis. The fourth data set D contains data of 28 subjects, that are similar to data set A. Instead of a goalkeeper perspective, subjects from study D were virtually placed in the center of a soccer field and had to remain overview over 360° in virtual reality. Data set A and D are combined to A* since it contains data of highly similar tasks. Each of the data sets contains subjects that were defined as experts (based on years of experience or being picked by talent scout), novices (beginners in the field or no experience in the task), and intermediates (loosely defined as in between, with more experience than novices but way less than experts). As not every data set was captured with the same model of eye tracker nor vendor, all the features from all data sets are looked at and a subset of features that are shared by all of the data sets is defined. In the next step, the data is split to experts, intermediates, and novices in each data set. A balanced training set of a randomly picked data set $x \in \{A^*, B, C\}$ is defined and a bagged tree model is trained. In this first model, the feature selection from Section 3.1.1 is used and the features are ranked by their importance for the model during the cross-validation. With this new subset of features data sets $y \in \{A^*, B, C\} \setminus x$ are used as the test set.

Results

Classification performance

After the feature selection process, the following features remain that have the highest impact on the classification and are therefore picked as candidates for the subset of features that might be shared by all the data sets.

- maximum saccade peak velocity
- maximum fixation dispersion
- standard deviation of saccade peak velocity
- maximum saccade amplitude
- minimum smooth pursuit dispersion

With the mentioned features an accuracy performance of 58% was achieved. This sounds quite low, but this is a three-class problem. Thus, the chance level of picking the right class is 33.33%. With 58% the accuracy is slightly worse than the doubled chance level. An accuracy of over 66% would lead to the fact, that single samples might be classified incorrectly, but the majority is classified correctly. Therefore, also the majority of a subject's samples are classified correctly and subsequently the subject in total, too.

Looking at the two data sets that were classified, the dentists' data set had a total classification accuracy of only 29%. The intermediates were classified with 7.7%, the experts with 35%, and the novices with 45 %. Thus, the dentists' data set is slightly worse than the chance level, and therefore, the most optimal features for that data set might not be found. Another reason for this classification might also be the totally different task of static diagnostics. There were restrictions on head movements as this study used a remote eye tracker while stimuli had been shown on a screen. On the soccer data set, which task was much more similar to the surgeons, accuracy for the novices reached 83.4%, 0.5% for the intermediates, and 82% for the experts. Again, because of the misclassifications of the intermediates, the average accuracy is at 60%. All in all, with the mentioned features a model is trained with one data set and the two other data sets are classified with an accuracy of 58%. On a deeper look at the single classes on the combined test set (soccer and dentists), one can see that the novices were nearly optimally detected (92%), the intermediates with 3.2 % not at all, and the experts still with an accuracy of over 79%. From 100 runs 34,700 samples were correctly classified as novice and 3,000 incorrectly as an expert. This is no problem, as it is known in expertise research there are subjects acting better than their initial classification. More problematic is the amount of samples that belong to the expert class but is classified as novice or intermediate. In 100 runs 30,000 samples were classified correctly as expert samples. 4,300 samples incorrectly as a novice, and 3,400 samples as intermediate. At first, these results look complex to understand, but a closer look at how the intermediates are defined reveals the ambivalence of these results and a weak point in the classification. This will be discussed in the corresponding discussion section of this paper (Section 4.2.1).

Shared, latent expertise features

Having a deeper look at the features and their characteristics, one can see three important correlations. As the data was normalized based on each data set individually, the values can be positive as well as negative. For comparison, this is important, as the correlations are only visible there. The surgeons' experts e.g. have a maximum saccade peak velocity of -221.560 $^\circ/s$, followed by the intermediates with -0.7267 $^\circ/s$ and the novices with 222.287$^\circ/s$. Comparing the values with those of the soccer players, one can see that the experts also have a highly

negative value of -593.31 $^\circ/s$ followed by a high value of 234.56 $^\circ/s$ and an even higher value of 1211.377 $^\circ/s$ for the novices. Soccer players show more or less the same trend between the expertise classes. In the data set of the dentists, this trend is not visible. Only experts and novices show similar values, thus, intermediates will be misclassified as novices (their values correspond much closer to the novices). For the dentists a correlation between the trends of the standard deviation of the peak velocity was found. The dentists as well as, the surgeons follow the same trend (experts: ca. -15°$/s$, intermediates: ca. 6.5 $^\circ/s$, and novices: ca. 10°$/s$). Here the data of the soccer players do not fit at all. A feature whose values correlate with both other data sets' experts and novices, is the minimum smooth pursuit dispersion. The values for the expert groups are slightly positive (0.019 to 2.25 pixels), while the values of the novices are slightly negative (-4.8 to -0.15). Only, again, the soccer players' intermediates correlate with the surgeons by being negatively close to zero.

3.2.2. Expertise Classification of Soccer Goalkeepers in Highly-Dynamic Decision-Tasks: A Deep-Learning Approach for Temporal and Spatial Feature Recognition of Fixation Image Patch Sequences

B. W. Hosp, F. Schultz, E. Kasneci, and O. Höner. "Expertise classification of soccer goalkeepers in highly dynamic decision tasks: A deep learning approach for temporal and spatial feature recognition of fixation image patch sequences," Frontiers in Sports and Active Living, vol. 3, p. 183, 2021.

Motivation

Although recent research focuses on behavioral features, there is a lack of understanding of the underlying cognitive mechanisms. First and foremost because of missing adequate methods for the analysis of complex and high-speed eye-tracking data that go beyond accumulated fixations and saccades. The latest research signifies that, until now, there is no feature set that allows general statements to be made, not even in the same domain. In fact, if there is no manually picked superior feature set that yields high-performance results, a rational step would be to use a learning algorithm to find features automatically, without a human in the loop. Hence, we investigate a different way of spatial and temporal feature recognition by using fixation image patch sequences. This approach removes arbitrariness and manual feature selection totally out of the process of defining predictor variables as the foundation of classification. A comparison of the automated feature selection versus a manual feature selection is done in the discussion of this work (Section 4.2.2).

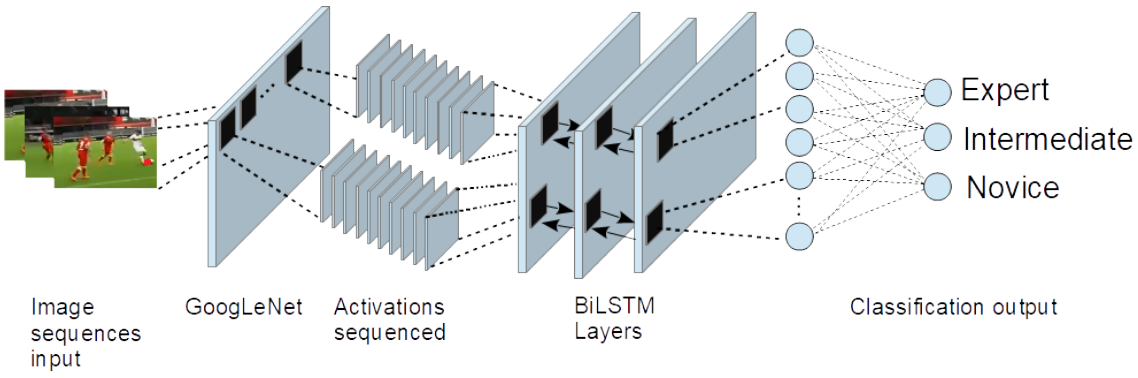

Figure 3.6.: Pipeline model of the classification network.

Methods

Our method includes the finding of latent features (hidden in the image patches gazed at during fixations) and subsequent classification of these patches as consecutive sequences. The aim is to predict three different expertise classes. For that, the same data as in Section 3.1.1 and all the fixation data is used to cut out the image patches, the subjects were gazing at during a fixation. For training, the images were augmented in several steps to adapt training to a realistic range of samples. The normal and the augmented samples are fed to a CNN (GoogLeNet) to find latent spatial features in the fixation patches. The procedure is illustrated in Fig. 3.6. To be able to do this, transfer learning is done on the GoogLeNet network, as this network is trained to recognize over 1000 classes of objects. The last layers after the last pooling layer is removed and added our BiLSTM network as well as a final three-class classifier to it. In each run 70% of the data belong to the training set and 30% to the validation set. All data of one subject is totally held out (hold-out-validation) to test our model with unseen data. One sample represents one video trial of one subject. Thus, the classification of expertise of subjects is looked at indirectly, instead, each sample is focused at. This means that some samples of the same subject can be detected to belong to different expertise classes.

Results

The average classification accuracy reached 73.11% over 33 runs. The accuracy of predicting a novice correctly as a novice is at 55.5%. While only 166 samples out of 1,816 samples in total were classified as expert samples, 650 novice samples were classified as intermediate samples. For the intermediates, there is a similar data situation. Out of 1,605 samples 119 were classified as expert samples and 372 as novice samples. 1,114 samples were correctly classified as intermediate samples, which corresponds to an accuracy of 69.4%. The best recognizable group is the experts. With 15 samples being classified as novices and 30 samples as intermediate, a very large majority is classified correctly as expert samples. The average accuracy of detecting expert samples correctly peaks at 93.4%.

3.2.3. Conclusion

The previously presented papers showed two different views on generalization. The first research question asked was whether one can find a subset of features that elicits expertise in a domain-independent manner. With the three data sets A* (containing gaze behavior of soccer goalkeeper and field player perspective in omnidirectional videos), B (containing gaze behavior of surgeons during a live arthroscopy), and C (containing gaze behavior of dentists during OPT analysis in 2D), indications were found that the similarity of the task seems to be important.

Our investigations showed that data of a static visual search task can not be properly classified, as where the classification of data from a highly similar dynamic task shows much higher classification rates. The surgeon and the soccer data set had much more in common. Dynamics (video), kind of eye tracker (able to look freely around or limited by a screen), and task (find a proper way to continue) are the most important. Therefore, with the current data situation, perceptual expertise cannot be stated to be domain-independent. The results only allow stating that expertise might be domain-independent as long as the task is familiar. This is well in line with [72], who showed that it is possible to transfer perceptual expertise from familiar tasks to semi-familiar tasks but not too unfamiliar tasks.

Regarding the data of the two familiar tasks, one can see that there are strong correlations. Obvious trends across the classes for both data sets are the smallest maximum peak velocities of saccades of the experts, followed by intermediates, and then novices with the highest value. For the unfamiliar task (dentist data set), one sees correlations between the experts and the novices, in the maximum peak velocity of the saccades, as well as in the standard deviation of the saccade peak velocity, but they were not strong enough to build a robust classification. Regarding a correlation between all data sets, one can see that the minimum smooth pursuit dispersion might be an indication that perceptual expertise can be domain-independent.

Our second research question was to investigate whether one can remove arbitrariness out of the way of defining features by focusing on simple features like image patches. It can be stated that deep learning, especially the combination of a CNN and an LSTM network is able to define and find spatial and temporal features independently, without the need for manual work. Our approach on the classification of fixation image patches shows high accuracy values, that can compete with traditional machine learning methods. The results might also be improved by adding more data on more subjects.

3.3. Gaze-Based Assistance Timing

3.3.1. States of Confusion: Eye and Head Tracking Reveal Surgeons' Confusion during Arthroscopic Surgery

B. W. Hosp, M.S. Yin, P. Haddawy, R. Watcharopas, P. Sa-ngasoongsong, E. Kasneci. "States of Confusion: Eye and Head Tracking Reveal Surgeons' Confusion during Arthroscopic Surgery." In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA.

Motivation

The use of eye-tracking methods to detect other cognitive processes besides expertise and cognitive load may be of particular importance for training systems that reduce training time. Online detection of confusion can contribute greatly to this. This is because if confusion can be detected in real-time during a task, then based on this, targeted assistance options can be applied to facilitate or even enable the continuation of the task. Targeted temporal detection of confusion can thus provide a basis for a training system. In particular, in arthroscopy, distracting information should not be shown on the screen during normal operation. Thus, a precisely timed view of auxiliary possibilities is directional with respect to perceptual-cognitive training systems. Young surgeons can thus shorten their training, which until now always required a supervisor to provide assistance in case of confusion. Optimal recognition of the right time for assistance can thus, on the one hand, accelerate and improve training and, on the other hand, save hospital resources through digital possibilities. To find an optimal timing for supportive measures, one has to focus on the detection of confusion. Confusion needs to be detected quickly and relatively accurately to signal the time when supportive actions need to be taken to assist non-experts in their learning process. The following work focuses on an online, machine learning-based approach to confusion detection in surgery. The findings of this work can be used to create training scenarios that not only optimize training for novices but also reduce the number of training hours that an expert must instruct. In addition to confusion detection, our goal is to create a fast model that can be used online. We want to predict confusion in real-time using a minimal set of features obtained from an eye tracker during surgery with an arthroscope. A long-term goal is to apply this method in an intelligent training environment that provides optimally timed assistance through temporally and spatially placed visual cues.

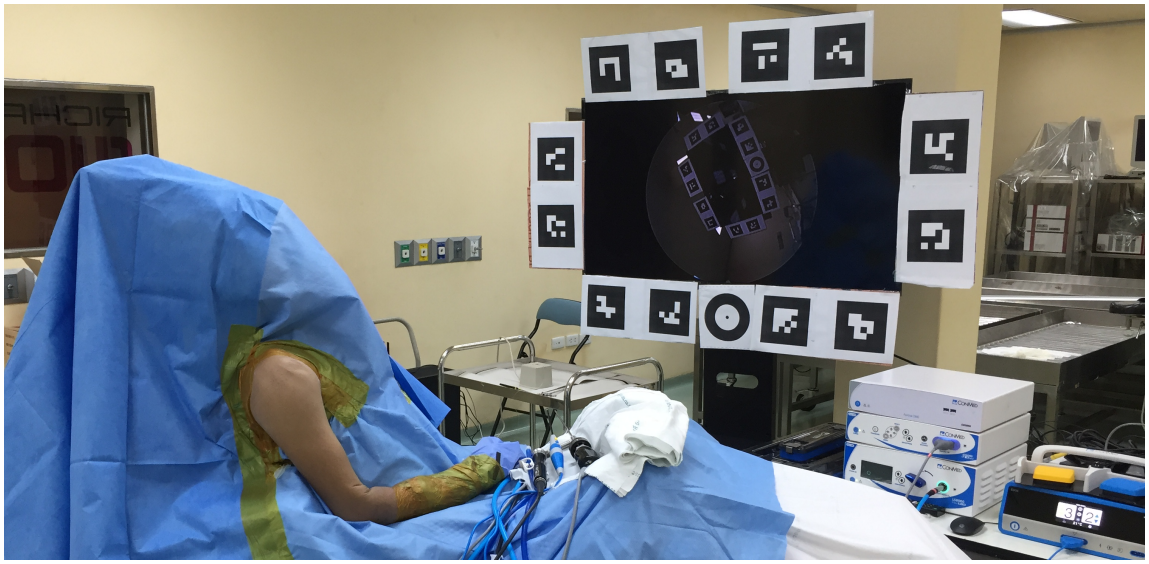

Figure 3.7.: Side view of the operation side with covered cadaver (left) and arthroscope output screen (right).

Methods

For the purpose of confusion detection, the data from Section 3.1.2 were used. There are have six novices that reported confusion during the surgery. For each of the moments, the samples from one second before until one second after the reported event were picked and were manually labeled as "confusion event". All other samples were labeled as "no event". Each sample contained the following features:

- point of regard (x, y)
- pupil position (average of both eyes)
- gyroscope (x, y, z)
- accelerometer (x, y, z)

To build a random forest model, the samples are split into training and testing data set. Out of 1,266,758 samples, there were had 7,103 samples with a confusion event and 1,259,655 samples with no event. Out of these samples, 7,103 confusion samples and 7,103 no confusion samples were collected. In every run, 1,000 samples of both were randomly picked to predict their class. The other samples were used for training. 2/3 of the subjects were randomly picked for training and counted the number of confusion event samples for each. Afterward, the same amount of "no event" samples from the same subjects was collected. This means

for our training set there was the same amount of confusion event samples as no event samples. This leads to a balanced training set (50% confusion event samples and 50% no event samples) and to a random baseline classification accuracy of 50%, which allows for easy interpretation of the results later. As the model is planned to be used in an online fashion, the classification accuracy (with unseen data) needs to be tested, as well as the classification speed. For that, a queue of n=2,000 samples is created. As the system was developed on saved data, the data were constantly read row by row. In each moment, there were n =2,000 samples in the queue, which represent one sequence. Every time a new sample is added to the queue, the oldest sample gets kicked out. Subsequently, the average of the features of the samples inside the queue is computed. These values represent the current content of the queue, which is called delta sample. This delta sample is now given to the trained random forest model to classify it as a "confusion sample" or a "no confusion sample". To infer the average performance time, the computation time of 100 single runs is measured and the average performance time is calculated.

Results

The average accuracy of the random forest model is 94.2%. According to the accuracy, the average misclassification cost/loss is 0.088. The optimal loss value for the test approach is reached at 49 trees with a misclassification cost of 0.085. In total, there are 50,000 samples for each class. Of class 0 (no event), 47,016 samples (93.8%) out of 50,136 samples were predicted correctly and 3,120 (6.2%) incorrectly. Similarly, for class 1 (confusion event), the model predicted 47,023 (94.3%) samples correctly as confusion events and 2,841 (5.7%) incorrectly. To measure the performance speed of the model, the computing time of each of the 100 runs is measured. On average each prediction takes 0.039 seconds. This corresponds to a frame rate of 25 fps.

3.3.2. Conclusion

With the available data, one is able to reach high accuracy on the detection rate of over 94%. This value and the sufficient speed of 39 ms are high enough to use this model in an online training system for young surgeons. In the next step, these assistive options need to be evaluated. Arthroscopy is one of the tasks that could benefit a lot from such a training system. In arthroscopy perceptual problems occur often, that are hard to explain verbally but are easy to solve digitally on the output screen. With this online system, these digital solutions can be applied directly. As such, there is no need for a supervisor to stand next to the trainee, as a correctly timed detection can remedy.

4. Discussions & Outlook

This chapter deals with the discussion of the findings of the papers discussed in this work with regard to application in practice and generalizability. To allow to expand a diagnostic system to an application and thus as a foundation for a training system in practice that is able to improve the perceptional skills of a subject correctly, there are several important aspects that have to be considered beforehand. A diagnostic model must not only correctly classify the right skill level of the subjects, much more important is a low false-negative rate. In the case of low-ranked subjects, those will become better at some point and therefore, rank higher as before. Thus, the diagnostic, which is necessary after each training run, needs to be sensitive to such changes in expertise. However, experts that have been classified as experts already, usually stay in the expert class. Sometimes even experts perform worse than their class average, but this is assumed to happen rarely. Thus, the false-negative rate (expert being classified lower as intermediate or novice) is an important metric that needs to stay low and has to be observed. The main aspects of a diagnostic system are:

- **the classification accuracy is as high as possible.** Based on findings from statistics, a system that has a prediction rate of 100% is not possible, as up to a point, human gaze behavior underlies idiosyncrasy [33], which implies a certain degree of difference. Therefore, the aim is to get as high as possible. This will only be able at a certain point where most of the variations within an expertise class have already been fed to the training of the model. Otherwise, the diagnostic system might fail to at least some new data. However, there is no security as when this point is reached, as even a human classification cannot guarantee 100% accuracy. Therefore, even after a long time of testing, the system will still be at a certain risk of misclassification which is tolerable but needs to be kept in mind.

- **false-negative rate needs to stay low.** This is much more possible and is based on the sharpness of the decision boundary between the classes, at least up to a certain degree. It is unlikely that every subject shows consistently high performance. In fact, it is assumed that there are some false-negative predictions. However, a proper diagnostic system can bypass the problem of rare false negatives by simply averaging the results of all runs for one subject.

Thus, a system needs to tell the current classification of the current training run (the current stimulus) and averaged metric overall training runs (all stimuli), which whitewashes some rare false-negative results, that describe the total performance of a subject better.

- **continuous re-training of the model.** After a diagnostic run, the data of such need to be fed to the diagnostic model to help to improve the sensitivity of the decision-boundary. This shrinks the risk of misclassifications very much.

The classification accuracy, as well as the false-negative rate, can already be described with the current state of the diagnostic models addressed in this work. The aspect of continuous re-training is not, as for this the system needs to be used frequently. As such, it is the task of future users. Therefore, the presented papers will be discussed on the first two points. As soon as these aspects are sufficiently addressed, the next step will be to define support options.

4.1. Gaze-Based Expertise

4.1.1. Soccer

The results of the expertise detection in soccer goalkeepers lead to several conclusions. Firstly reached several milestones from the aforementioned aspects of a diagnostic system were reached, and secondly the process of understanding what it takes to perform well was simplified. Algorithmically the process of feature selection was improved, which is important at the beginning of every classification procedure. One important note is that it is vital to assign samples in a person-specific manner. Idiosyncrasy is a dead end, as it suggests high-performance classification, but fails completely in the prediction of new unknown samples. This is fundamentally based on the relationships of the differences between different groups. Samples of the same person have the highest similarity (idiosyncrasy). Samples from the same group come next as the differences across experts are shown to be smaller than the differences across groups of expertise (expert variation). The accuracy of the detection rate allows - with over 78% - an application in practice. There is still some scope for optimization, but at its current state, the model is already quite performant. It is common in expertise research that subjects, assumed to rank low, might surprise with high talent and therefore get ranked in higher classes. Foremost, this can be counteracted by relying on strict rules when defining each of the classes. This is especially true for the novice class. Since our assignment is based on a relative rule (no experience in competitions and no training on a regular basis), which allows the feature characteristics of the novices to spread on a wide area but lead to strict decision boundaries for the intermediates and experts.

A strict differentiation within the novices might reveal further differences in gaze behavior between novices with no experience at all and novices with experience being a long time ago. With the current status, the expertise of novices has never been assigned professionally. Thus, the recognition of the novice group might be harder than the alternatives. This is especially evident in the false-negative and false-positive rates. 18.6% of the novice samples were classified as intermediate samples, but only 1.6% of the intermediate samples as novice samples. A portion of low performers usually can be found in higher-performing classes. And as there is no ground truth for novice expertise, the classification of a novice is considerably more difficult. A much more important finding is, that our model fundamentally has an extremely low false-negative rate, thus, only a few subjects are wrongly switched from higher to lower expertise classes. Which is one of the main aspects of a diagnostic system.

Further, for a reproducible and objective way, the way how features get selected as predictor variables is essential. Features can have different influences on accuracy. Feature selection is done in order to optimize the accuracy of the predictions and by selecting only a subset of features, a dimensionality reduction is achieved, too. With less dimensionality, the problem is becoming less complex, and computational power and speed can be saved. But the methods of feature selection can have other purposes, too. A reduced feature set can help to avoid overfitting of the model. Using fewer features reduces the risk of the model memorizing certain training examples. Likewise, fewer features can improve the interpretability of the model, as the affection of certain features can be identified. Thus, our model reached another milestone: real-time operability. With a low number of features, the computation time is held low, which speeds up the whole process. For the use as a diagnostic system, we assume that only the classification accuracy needs to be improved further. This can be achieved by collecting more data from more subjects to robust the model against outliers.

The simplification of the understanding of gaze behavior is based on the average characteristics of latent gaze features that represent each of the expertise classes. Novices tend to do a lot of small movements with their eyes. We interpret that behavior as a signal of tension or nervousness, as they have less experience and try to perceive every possible change in the scene around them as fast as possible. These results are not final but can be used to teach a more planned scanning behavior to novices. The characteristics might even change as soon as there is enough data collected from further subjects, because the lower the number of subjects, the more outliers make a difference.

In summary, the current state of the model allows the usage as a diagnostic model in practice, because both aspects have been addressed sufficiently. The accuracy will usually grow higher as soon as more data is available. Thus, continuous re-training will considerably be important. For the use in a training system, one first idea is to use the general rules from the physical training of the specific do-

main. At different states of expertise, there are different aspects of perception that are being trained. This can be the perfect timing for shoulder glances, finding and utilization of free spaces, or just an optimal decision after a pass.

As the whole system is data-driven and only based on the gaze signal, the model can be used in other domains, too. The way how features are calculated and evaluated to be part of the classification, as well as the classification process is highly automated and as such, the pipeline to build the model provides high generalizability. The only requirement is that the gaze signal can be obtained online or at least from a file. Further details about the generalizability of this approach can be found in the work in Section B.1.4 and Section 4.2.1 where the same approach is applied to data sets from different domains.

4.1.2. Orthopedics

In the field of arthroscopic surgery in orthopedics, several indicators were found that allow the separation of experts, intermediates, and novices. First of all, novices need much more time (up to two times) to solve the same tasks as experts. This is also reflected in their gaze behavior. During the initial search phase where subjects navigate to the main area where the next landmark is placed, novices were acting the same as experts. The difference is in the last 20% of the search, where novices had problems in fine-tuning the arthroscopic camera to the correct place. Thus, they needed more time to reach the landmark. Novices were also the only group that reported confusion. During such a state of confusion their cognitive load grew, which could be shown with the PCPD of about 1% compared to baseline. The typical metrics of gaze behavior could only lead to an understanding of the differences between the experts and the novices, but not any further. Since a three-class classification is considerably more complex, investigations in another way of feature selection are needed. With the use of MRMR methods and chi-square-test from Section 3.1.1, differences between the three groups were found and at least 76.46% of the differences could be explained.

Regarding the two points raised earlier about the application of the model in practice, the model is quite powerful, but to improve the training time of young surgeons the performance needs to be improved further. A miss rate of 23.54 % is still too high for an application in practice. On the plus side, this is only based on 15 subjects, which means there might be even other feature combinations that explain differences much better, but for a pilot study, the values are quite promising. To continue further, there needs to be more data collected from more subjects. A finer graded classification would help to understand the differences even further. With this study, the same challenges as with the soccer goalkeepers study are present, but as there is a really low number of subjects, fewer aspects could have been addressed. The only milestones that were reached with the current state of this

work are to prove that there are differences between the classes, which can be found by looking at their gaze behavior, and a fast computable subset of features, which allows an online application of the system.

Luckily, the four features are easily calculated, which would allow the usage of the classification as an online classification system. Though, the classification needs to be done segment-wise after a certain period of time, which needs to be investigated first. Further steps are to add more subjects to each class and refine the number of classes. This allows a much finer classification and therefore a better understanding of the differences between the classes. A finer classification is important to robust assumptions made by the model about gaze behavior and optimizes the recognition of class-specific weak spots to be used in a training system.

A high miss rate of 23.54% symbolizes also a high false-negative rate, which would violate the second requirement for the application in practice. Similar to the model in Section 4.1.1, continuous re-training of the model with much more subjects is essential here. However, until more data is available, no robust statement can be made about the application in practice. Again, the procedure to develop this model is highly generalizable as the same method are used for feature calculation, as well as feature selection and model creation as in the model of Section 4.1.1. As far as generalizability is concerned, the same conclusion can be drawn as in Section 4.1.1. However, due to the typical character of a pilot study which is based on a small number of subjects, there are limits to the informative value of the features used in this model in terms of their generalizability to a larger data set from the same domain and task. The features may change completely as more data is available.

4.1.3. Outlook

Although only one model from Section 3.1.1 meets the requirements for use as a diagnostic system, the findings of both papers are far-reaching. Indeed, it can be said that eye tracking is in many ways well suited not only to detect perceptual differences but can even provide more profound information that different classes can be inferred from different gaze behaviors. Eye tracking is thus an essential part in the study of human behavior and will very likely continue to be able to provide important insights from the aspect of multimodal interaction between humans and machines. As mentioned before, one of these aspects could be a training system based entirely on eye tracking. Now, on the one hand, the models have to be made more robust by adding more data to learn from, and, on the other hand, the first steps have to be taken into the application. Since especially in the fields of soccer and surgery new ideas for improving training are constantly sought, such training systems meet a relatively large market, as soccer clubs and hospitals are

increasingly turning towards digital possibilities. Also in view of current trends in human-computer interaction, which aim at a personalizable self-diagnosis, this system would be an important building block, as the procedure performed in the two sections can be applied to other perceptual-cognitive processes as well since the procedure works independently of the training data labeling. To adapt the model to another perceptual-cognitive process, only the training data must be marked correctly.

4.2. Cross-Domain Generalization

The understanding of expertise is mostly limited to a certain domain or task. Thus, it is assumed that perceptional expertise is it, too. However, what is if this is not the case? If perceptional expertise is something everyone can learn, cross-domain training systems can be build that help subjects to improve in several domains simultaneously. An ophthalmologist can provide perceptional expertise tests. They might be able to diagnose different levels of perceptional expertise and thus, the aptitude of a subject for a certain task or even diagnose diseases that hinder one to apply an optimal gaze behavior. Simple tests can be provided that can be included in several fields where perception plays a central role. Also, our understanding of perception might change completely. As visual perception can play a decisive role in soccer and medicine, highly specified training should become an essential part of education. The first steps in the direction of general perceptional expertise detection are discussed in the following.

4.2.1. Expertise-Related Features

One of these first steps is the search for shared features that describe the same expertise levels of subjects from different domains. With a model that is able to find commonalities between data sets from different domains, and thus, differences between expertise classes that are valid across domains, there is an important starting point for the search for generalized expertise.

When looking at the single values of the features used of the different classes and domains, it can be stated that there are correlations between the classes and between a subset of domains. Correlations have been found in the maximum saccade peak velocity, the standard deviation of the peak velocity, and minimum smooth pursuit dispersion between subsets of the domains included. As not all domains share the same features equally, generalizability is hard to state. To answer the question about generalizability, first, the model needs to be trained with more data of the known domains (but equally from each class) and of foreign domains. However, with the current work, it is possible to find commonalities between different

domains that reflect expertise in a general way. The findings suggest that it is worth continuing the search for general expertise features. There are two optional ways to go now. If the search for commonalities between domains is considered as a ternary problem, the detection of the single classes, especially the intermediates, needs to be optimized first, as it is close to 0%. Therefore, in its current state, the model can not be used to build an application. The findings are too weak to function as a basis. If only novices and experts are considered and intermediates are disregarded, the model is performing quite well already. Both expertise classes show accuracy values more than doubled chance-level. Thus, they are sufficient to classify the majority of samples of a subject correctly and therefore reach a low false-negative rate, too. So, by defining domain-independent visual expertise as a binary problem, the model can be used in an application to detect general visual expertise, but solely classified as novice or expert (If there is any use case where a binary differentiation is desirable).

However, in any case, it is astonishing that such fundamentally different domains and tasks have such a load of common perceptional features. In both cases, to robust the model, first of all, there need to be strict definitions of the single classes that allow a proper classification. In soccer, this can be the years of experience or hours practiced. It just has to be more precise than "has never participated in a competition". This definition is too vague. The classification result of 3.2% shows that intermediates from one data set are not equally skilled as intermediates of the other data sets. Thus, intermediates of one data set are considered as being better/worse and end up being classified as expert/novice. Only if the classes are clearly separable, the machine learning classification can perform well. Another step towards an application is to train the model with more data. This means more subjects from different expertise groups, but also more subjects in total. At the moment the classification accuracy is 58%, thus, slightly worse than double the chance level, which indicates there is still enough uncertainty that too many samples of a subject will get classified incorrectly. As soon as the model reaches over 66% it is strong enough to classify the majority of a subject correctly. This would be sufficient to continue the search on general expertise features. So far, it is only know that there are common perceptional features between surgeons doing an arthroscopic surgery, soccer players in decision-making situations, and hints about some commonalities to dentists on a visual search task. The next major step would be to investigate even more domains and see whether the features found in this work can be transferred to even more domains. Our findings suggest considering some aspects. The task and the requirements for the technology (same kind of eye tracker, similar dynamic scenes, etc.) should fit each other. We assume the better the technology and the scene and task fit together, there are more commonalities visible in the gaze signal.

4.2.2. Deep Learning Feature Selection

Another important step in the direction of generalized perceptional expertise is the way the data received are examined. So far, manual selection of features that are thought to be helpful is the main approach. Subsequently, machine learning techniques are more often used. One specific advantage is that a machine is not at risk of arbitrariness and as such, the generalizability of such techniques might be higher than with manual feature selection. Usually, machines select features because of their importance based on a calculation. Thus, removing the arbitrariness out of the process of feature selection.

In our current model, the accuracy of predicting an expert correctly is at 93.4% as this class is the easiest to detect. The prediction rate of the intermediate class is much lower with an accuracy of 69.4% because this class is supposed to be the hardest to detect. The accuracy, however, is more than double the chance level with about two-thirds of the intermediate samples being classified correctly. Much lower than the intermediates, the novices are predicted with an accuracy of 55.1% which is nearly two times as high as the chance level but still 11% lower. The expert group is a pretty well recognizable group. The intermediate and novice groups are more heterogeneous as there are subjects that have more/less experienced than others. Another reason for this could be the missing metrics needed to divide between the two classes properly. This question is typically addressed with the availability of more data. The problem may stem from the small sample size of intermediate subjects as this group could be too small for the model to define robust decision boundaries. The fact, that expert samples were barely (15 samples) predicted to be samples of an intermediate player, shows that there are clear decision boundaries for the intermediate and expert classes. Nevertheless, a long-term goal is to optimize the training for young players, whereas This study is the first step in that direction. For that, one needs to know which behavior is optimal and how one can design training steps for young players to reach this optimal behavior. The difference in active years/training, and therefore experience, between intermediate and expert subject, is much smaller and needs to be finer graded. As soon as the detection of the novices is more robust, the model is likely to become an application that is used in practice.

Especially instead of providing a description of the behavior of different classes, this model describes a pipeline to find latent features by itself. This circumvents one problem: handcrafted features. The characteristics of handcrafted features may be difficult to teach a user in the form of new behavior based on feature values. Even if the optimal set of features is found, it is difficult to incorporate the findings into a training system.

Conclusively, one can state that a certain degree of automation has been achieved in the process of feature selection. This improves the whole process as now fea-

tures are used that are not only said to be meaningful, but that can algorithmically be calculated and explained. As such, the model shows high generalizability, as an only requirement the scan path of the subjects (independent of the task) as fixation image patches needs to be provided. The remaining pipeline is automated totally. However, a comparison between the manual selection in Section 3.1.1 and the automatic one of this work, shows that manual selection performs superiorly. One explanation might be, that for the current state of the data set with a still-low number of subjects, the essence of what is important can not be robustly depicted in the fixation sequences. It might also be that the error of the eye tracker has a higher impact on the absolute values of the fixations than on the relative features picked in Section 3.1.1. At the moment, the manual selection is better suited, but sooner or later, at least when there is much more data available for training and eye tracker errors can be excluded, machine learning methods will pass manual selection. The correct data representation to feed them with are just not found.

4.2.3. Outlook

From these works, one learned that there are differences and commonalities in the classes across the data sets. Similar to the papers on classification one faces the same problem of defining expertise. Commonalities between experts and novices were found, which is a first step in the direction of understanding how expertise develops, but an understanding of intermediate subjects' visual perception lacks. Thus, one need to focus on strict definitions of each expertise class properly. The current definition is too vague. In fact, the cognitive factor is only one of the several factors that contribute to expertise. For goalkeepers, for example, it is still most important to be able to block shots on goal. If a goalkeeper can do this extremely well, they may be invited by the DFB and classified as an expert), even though they could make "worse" decisions after return passes. Conversely, it can also be the case that intermediate subjects are very good decision-makers, but did not hold as many balls, which is why they are not invited by the DFB. As a result, it is very important to not just test players from different classes but to test players with the assumed highest decision-making skills. The same situation exists in surgery and dentistry. The pre-classification as ground truth might classify a 4th-year resident as an intermediate because they are in the 4th year of the residency, but how much practice this surgeon has already had is not taken into account. In dentistry, the exact same problem is faced. As such, it is useful to search for commonalities and especially for general features of perceptional expertise, independent of the status of their education. By doing that, assistance options can be simply provided to everyone that wants to optimize their perception.

In fact, this model offers a different way of teaching a subject a new behavior by visualizing the test person what and when has to be observed. Therefore, a

model should be created that in the best case, finds an optimal behavior. Based on such information, an optimal behavior for each class can be created and artificially extracted to create information that can be taught to users. A prerequisite will be the analysis of single scan paths, which can be accessed by looking at the fixation image patches. Currently, as the fixation point is temporally and spatially averaged, another improvement might be achieved when optimizing the input layer by using an object detection beforehand. Especially when counting in the error rate of the eye tracker and early fixations, some samples might end up directly next to an object and some directly on it. In this case, the CNN will return different shapes. By using the object as an area of interest (AOI) and taking the intersection as input, this behavior can be unified as one can assume that the subject is perceiving the same object in both cases. The CNN can also be optimized. At the moment this CNN is trained on ImageNet to classify about 1,000 classes. By retraining the CNN on a set of 360° videos, with manually labeled teammates, opponents, goals, the ball, and free spaces, the intersections of the gaze with AOIs can become advantageous and result in higher classification rates.

4.3. Gaze-Based Assistance Timing

With detection of a gaze-based assistance timing, one can for example help surgeons to proceed, by either pointing out visual clues, which may be used by expert surgeons to navigate or drawing arrows on the output of the arthroscope which tells the surgeon where to navigate next (examples are shown in Fig. 4.1). Another possible usage of the knowledge of the correct timing for assistance can be to augment the whole output by describing the scene by segmenting and labeling each bone or tissue. Or simply name the shown parts in the output. There are multiple ways of supporting a surgeon. Depending on the state of expertise, the level of assistance may be chosen, to allow different skilled surgeons, to train their different weaknesses.

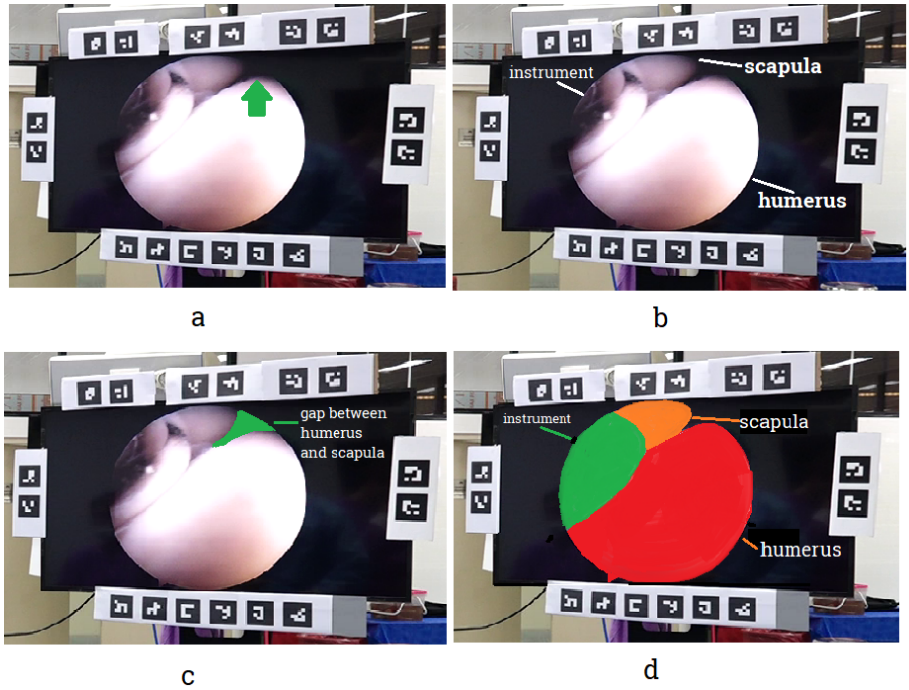

Figure 4.1.: Assistance options for young surgeons to regain overview.

4.3.1. States of confusion during arthroscopic surgery

To enable the use of the models presented so far in the form of a diagnostic application, another feature is required. Of course, assistance options can be constantly displayed to the user during training. However, this can lead to the user being overwhelmed by the number and frequency of the assistance options, so that they may distract from the actual training. To prevent this, it makes sense to use the methods presented in Section 3.3.1. Because by knowing the right time for assistance options, limited to short time intervals, exact help can be displayed, so that it will only appear when it is needed.

In this context, when a certain prediction rate is reached, the detection of optimal times enables much more fine-tuned and personalized assistance. Thus, personal weaknesses can be identified online and training can be tailored specifically to the current user. However, in order to incorporate such diagnostics into an application, two key aspects must be considered. First, the prediction rate must be sufficiently high. Thus the presented model is 94%, thus, more than adequate. And secondly, not only the right timing of the assistance but also an adequate form of it must be used. Because even if the right timing has been found, the right assistance must also be displayed to match the user's expertise class. Since different expertise classes can have different problems, these must be explicitly assigned in order to

display class-compliant assistance. Further, in this model, the false-positive rate (detection of confusion when none is actually present) should be kept low to avoid displaying unnecessary help, but the false-positive rate does not have a large impact on the applicability of the model, since it only indicates how often the user is shown additional information that may at most distract him a little because it was unnecessary for that moment. Besides the prediction rate, a short calculation time is also an advantage. Getting assistance only after 5 seconds is not efficient, because the scene may have already changed completely in this time step. Assistance options must be provided as quickly as possible.

Regarding the transferability of the model to other domains, there are no particular restrictions. If a think-aloud protocol is also used in that new domain, the presented methods can be used to train a model that detects and classifies states of confusion. To what extent the presented model trained on surgery data can classify new data from other domains is difficult to say. For this to work, the gaze feature expressions detected here during a state of confusion would have to be the same or at least similar to those during the confusion in the new domain. Probably, a direct transfer of the trained model without new training requires at least some similarities in the nature of the eye tracker and perhaps even the task. However, again, since there is little inter-task as well as inter-domain work, this needs to be found out in future work.

4.3.2. Outlook

One reasonable step to improve the model would first be to speed the whole pipeline up. That is, to enable the model to be used online by eye trackers with higher frequency. However, one essential step for this model is the design of adequate assistance options. For example, it does not make sense to label the arthroscopy image with the names of the bones and tissues if the surgeon is not yet familiar with them or currently learns how to navigate the arthroscope camera. Likewise, it is meaningful to specify the task beforehand and then purposefully train for one specific weakness. In football, for example, this can be done by configuring training to optimize shoulder glances. This way, the user knows what they have to pay attention to and the application can specifically display help options such as free spaces or trigger the correct direction of the gaze via sounds. There are a plethora of ways to assist young trainees, but which are reasonable, needs to be investigated in future works.

5. Ethical Considerations

In general, when recording, recognizing, and processing personal, biometric data, which may include eye-tracking data, special care must be taken, as such data require special protection. This is true not only from a legal point of view but especially from a research ethics point of view. All study data used in this work were collected in accordance with Article 4, No. 1. DSGVO with the consent of the subjects. Subjects have the right to view and delete their data at any time. Data that, in our view, have a direct personal reference, were stored anonymously and were only used for the work mentioned in this thesis in the sense of research. If data had to be published for publication, this was done in accordance with the applicable rules and laws. Likewise, where possible, all work was published under licenses prohibiting commercial use. Since in my view research is done for society.

In addition to the special nature of biometric data, research data on expertise recognition represent, in my view, another special feature. An identification of persons by their characteristics and the associated diagnostic data (expertise, confusion) was prevented by strict anonymization. This prevents any possible damage that could be caused by the identification of the test persons. Particularly with regard to artificial intelligence, there are often uncertainties about data protection, so I would like to make a special note here that all study data were only used anonymously so that neither outsiders nor algorithms (e.g. artificial intelligence) can establish connections to individual persons.

A. Gaze Expertise Linkage

This chapter is based on the following publications:

- **B. W. Hosp**, F. Schultz, O. Höner, and E. Kasneci. "Soccer Goalkeeper Expertise Identification Based on Eye Movements." In: PloS one, 16(5). 2021.

- **B.W. Hosp**, M.S. Yin, P. Haddawy, P. Sa-ngasoongsong, and E. Kasneci. "Differentiating Surgeons' Expertise Solely by Eye Movement Features". Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA.

- M. S. Yin, P. Haddawy, **B. W. Hosp**, P. Sa-ngasoongsong, T. Tanprathumwong, M. Sayo, and A. Supratak. "A Study of Expert/Novice Perception in Arthroscopic Shoulder Surgery". In Proceedings of the 4th International Conference on Medical and Health Informatics (pp. 71-77). August 2020.

A.1. Soccer Goalkeeper Expertise Detection Based on Eye Movements

Abstract

The latest research in expertise assessment of soccer players has affirmed the importance of perceptual skills (especially for decision making) by focusing either on high experimental control or on a realistic presentation. To assess the perceptual skills of athletes in an optimized manner, we captured omnidirectional in-field scenes and showed these to 12 experts (picked by DFB), 10 regional league intermediate players and 13 novice soccer goalkeepers on virtual reality glasses. All scenes were shown from the same natural goalkeeper perspective and ended after the return pass to the goalkeeper. Based on their gaze behavior, we classified their expertise with common machine learning techniques. Our results show that eye movements contain highly informative features and thus enable classification of goalkeepers between three stages of expertise, namely elite youth player, regional league player, and novice at high accuracy of 78.2%. This research underlines the importance of eye tracking and machine learning in perceptual expertise research and paves the way to perceptual-cognitive diagnosis as well as training systems.

A.1.1. Introduction

Along with physical performance factors, perceptual-cognitive skills play an increasingly important role as cognitive performance factors in sports games. In perceptual research examining the underlying processes of these skills, subjects are typically placed in a situation where they have to react while their behavior is being recorded and subsequently analyzed. Such behavior can be assigned to a class, for example, to provide information about performance levels. Many studies in sports games in general, and in soccer in particular, [61], [117]–[122] have shown that athletes in a high-performance class have a more highly developed perception, leading – amongst other factors – to success in sports. However, this research is still confronted with challenges regarding experimental control and a representative presentation of the situation. Furthermore, the potential of novel technologies such as eye tracking as a means to assess the underlying perceptual-cognitive processes has not yet been fully exploited, especially with regard to the analysis of complex eye-tracking data. In this work, we research how to handle and analyze such large and complex eye-tracking data in an optimized way by applying common supervised machine learning techniques to the gaze behavior of soccer goalkeepers during a decision-making task in build-up game situations presented as 360°-videos in a consumer-grade virtual reality headset.

Latest sports-scientific expertise research shows that experts - when it comes

to decision-making- have more efficient gaze behavior because they apply an advanced cue utilization to identify and interpret relevant cues [67]. This behavior enables experts to make more efficient decisions than non-experts, e.g. during game build-up by the goalkeeper. From both a scientific and practical sports perspective, of particular importance are factors that lead to successful perception, form expertise, and how these can be measured. To measure perception-based expertise, at first, a diagnostic system is needed for recognition of expertise, which provides well-founded information about the individual attributes of perception. These attributes are usually considered in isolation. Thus, their influence on expertise can be specifically recognized. To allow the athletes to apply their natural gaze behavior, the experimental environment is important, but one of the main problems in perceptual-cognitive research persists in realism vs. control. In a meta-review of more than 60 studies on natural gaze behavior from the last 40 years, Kredel et al. [63] postulate that the main challenges in perception research lie in a trade-off between experimental control and a realistic valid presentation. Diagnostic and training models are often implemented or supported by digital means.

This is nothing new, as in sports psychological research, new inventions in computer science such as presentation devices (i.e. CAVE [123], virtual reality (VR) [124]), interface devices (i.e. virtual reality, leap motion, etc.), or biometric feature recognition devices (i.e. eye tracker [125]) are used more and more often. As a new upcoming technology, virtual reality (VR) devices are used more frequently as stimulus presentation and interaction devices. As said, a fundamental aspect in perception research is a highly realistic presentation mode, which allows for natural gaze behavior during diagnostic. VR technology makes this possible by displaying realistic, immersive environments. However, this strength, allowing natural gaze behavior, comes less from the VR technology itself. According to Gray [126], the degree to which the perceptual-cognitive requirements of the real task are replicated in such environments depends on psychological fidelity. Next to immersion and presence, Harris et al. [127] suggest the expansion of a simulation characterization into a typology of fidelity (containing also psychological fidelity) to determine the realism of a simulation. VR offers an immersive experience through the use of 4k 360°video, which experiences a higher level of realism than, for example, CAVE systems, by providing higher levels of psychological fidelity [126], [127]. VR is therefore a popular and optimal tool for perception research. Bideau et al. [128] summarize further advantages of VR in their work. Their main contribution, however, is their immersive virtual reality that elicits expert responses similar to real-world responses.

In a narrower sense, VR is based on computer-generated imagery (CGI). One advantage of such fully CGI-based environments is the possibility of the user interacting with the environment, which presumingly increases the immersive experience. On the other hand, fully CGI-based environments contain moving avatars that are natural in appearance and hide environmental influences. This might pre-

vent high immersion and influence the participant's gaze behavior. Therefore, we chose a realistic environment with 360°stimuli to provide a close to a natural environment that does not influence the participant's gaze behavior. As this work presents a focus on the cognitive processes of decision-making, we focus less on realistic interaction methods

Especially interesting are the developments of VR devices regarding integrated measuring devices. More and more devices have eye trackers directly integrated, which, in combination with a photo-realistic environment in VR glasses, allows for the measurement of almost optimal user gaze behavior while also showing highly realistic stimuli. Eye trackers provide a sound foundation with a high temporal and spatial resolution to research perceptual processes. The combination of VR and high-speed eye tracking allows the collection of a massive amount and highly complex data. With the high-quality eye images and freedom of movement of a mobile eye tracker, the high speed of a remote eye tracker and the control over the stimulus in a lab setting (VR), and the naturality of in-situ stimuli by omnidirectional videos, the outcome of this combination is highly complex. Analysis of such data is a particular challenge, which emphasizes the need for new analysis methods. As we want to infer underlying mechanisms of perceptual-cognitive expertise, tracking eye movements is our method of choice in this work. Generally, perceptual research focuses on eye tracking because, as a direct measuring method, it allows for a high degree of experimental control. Besides a realistic presentation and high degree of experimental control, VR can also be used to model the perception [129] of athletes and thus creates a diagnostic system. A diagnostic system has the ability to infer the current performance status of athletes to identify performance-limiting deficits, an interesting provision of insight for the athletes and coach as well. Most importantly, such a diagnostic system forms the basis for an adaptive, personalized, and perceptual-cognitive training system to work on the reduction of these deficits.

So far, eye-tracking studies have focused on either in-situ setups with realistic presentation mode and mobile eye trackers (field camera showing the field of view of the user) or on laboratory setups with high experimental control using remote eye trackers [130]–[135]. Since mobile eye trackers are rarely faster than 100-120 Hz because saccades and smooth pursuits cannot be detected properly at such speed, investigations in an in-situ context are limited to the observation of fixations. Fixations are eye movement events during which the eye is focused on an object for a certain period of time (and thus projects the object onto the fovea of the eye) so that information about the object can be cognitively processed. The calculation of fixations with such a slow input signal leads to inaccuracies in the recognition of the start and end of the fixation. Only limited knowledge can be gained using such eye trackers because additional information contained in other eye events, such as saccades and smooth pursuits, cannot be computed correctly. This prevents the use of such eye trackers as robust expert measurement devices.

Saccades are the jumps between the fixations that allow the eye to realign. They can be as fast as 500 °/s. Smooth pursuits are especially interesting in ball sports because they are fixations on moving objects i.e. moving players. However, especially in perception studies in soccer in VR-like environments, slow eye trackers with about 25-50 Hz are primarily used [136]–[139]. This speed limits the significance of these studies to fixation and attention distribution in areas of interest (AOI). Aksum et al. [138], for example, used the Tobii Pro Glasses 2 with a field camera set to 25 Hz. Therefore, only fixations or low-speed information is available and no equal stimuli for comparable results between participants. In a review of 38 studies, McGuckian et al. [62] summarized the eye movement feature types used to quantify visual perception and exploration behavior of soccer players. Except for Bishop et al. [140], all studies were restricted to fixations thus restricting the gainable knowledge of eye movement features. The integration of high-speed eye trackers into VR glasses combines both strengths: high experimental control of a high-speed eye tracker and a photo-realistic stereoscopic VR environment.

With more frequent use of eye trackers, and more accurate, faster, and ubiquitous devices, huge amounts of precise data from fixations, saccades, and smooth pursuits can be generated which cannot be handled in entirety utilizing previous analysis strategies. Machine learning provides the power to deal with huge amounts of data. In fact, machine learning algorithms typically improve with more data and allow - by publishing the model's parameter set - fast, precise, and objective reproducible ways to conduct data analysis. Machine learning methods have already been successfully applied in several eye-tracking studies. Expertise classification problems in particular, can be solved as shown by Castner et al. in dentistry education [29], [30] and Eivazi et al. in microsurgery [22], [97]–[99]. Machine learning techniques are the current state-of-the-art for expertise identification and classification. Both supervised learning algorithms [29], [97] and unsupervised methods or deep neural networks [30] have shown their power for this kind of problem-solving. This combination of eye tracking and machine learning is especially well suited when it comes to subconscious behavior like eye movements features as these methods have the potential to greatly benefit the discovery of different latent features of gaze behavior and their importance and relation to expertise classification.

In this work, we present a model for the recognition of soccer goalkeepers' expertise in regard to decision-making skills in build-up situations by means of machine learning algorithms relying solely on eye movements. We also present an investigation of the influences of single features on explainable differences between single classes. This pilot study is meant to be the first step towards a perceptual-cognitive diagnostic system and a perceptual-cognitive virtual reality training system, respectively.

A.1.2. Methods

The basis of this work is a pilot study on a VR system with an integrated eye tracker. This chapter describes the experimental setup, the pilot study, the eye-tracking characteristics, and the methodical procedure for the analysis with machine learning methods.

Experimental setup

In this study, we employed an HTC Vive, a consumer-grade virtual reality (VR) headset. Gaze was recorded through the integration of the SMI high-speed eye tracker at 250 Hz. The SteamVR framework is open-source software that interfaces common real-time game engines with VR glasses to display custom virtual environments. We projected omnidirectional 4k footage on the inside of a sphere that envelopes the user's field of view, which leads to high immersion in a realistic scene.

Stimulus material

We captured the 360°-footage by placing an Insta Pro 360 (360°camera) on the soccer field on the position of the goalkeeper. Members of a German First League's elite youth academy were playing 26 different 6 (5 field players + goalkeeper) versus 5 match scenes on one half of a soccer field. Each scene was developed with a training staff team of the German Football Association (DFB) and each decision was ranked by this team. There were 5 options (teammates) plus one "emergency" option (kick out). For choosing the option rated as the best option by the staff team, the participant earned 1 point, because this option is the best option to ensure the continuation of the game. All other options were rated with 0 points. Conceptually, all videos had the following content: The video starts with a pass by the goalkeeper to one of the teammates. The team passes the ball a few times until the goalkeeper (camera position) receives the last return pass. The video stops after this last pass and a black screen is presented. The participant now has 1.5 seconds time to report which option they've decided on and the color of the ball which was printed on the last return pass (to force all participants to recognize the last return pass realistically).

Participants

We collected data from 12 German expert youth soccer goalkeepers (U-15 to U-21) during two youth elite goalkeeper camps. The data from 10 intermediates was captured in our laboratory and comes from regional league soccer goalkeepers (semi-professional). Data from 13 novices came from players with up to 2 years

of experience with no participation in competitions and no training on a weekly basis. The experts have 8.83 hours of training each week and are 16.6 years old on average. They actively played soccer for about 9 years, which is significantly more than the novices (1.78 years), but less than the intermediates (15.5 years). This may be a result of their age difference. The intermediates are 22 years old on average but have nearly half of the training hours per week compared to the experts. Characteristics of the participants can be seen in Table A.1.

Participants

Class	Attribute	Average	Std. Dev.
Experts	Age	16.60	1.54
	Active years	9.16	5.04
	Training hours/week	8.83	4.27
Intermediates	Age	22.00	3.72
	Active years	15.50	5.77
	Training hours/week	4.94	0.91
Novices	Age	28.64	3.72
	Active years	1.78	5.21
	Training hours/week	0.00	0.00

Table A.1.: Participants summary.

Procedure

The study was confirmed by the Faculty of Economics and Social Sciences Ethics Committee of the University of Tübingen. After signing a consent form to allow the usage of their data, we familiarized the participants with the footage.

The study contained two blocks consisting of the same 26 stimuli in each (conceptually as mentioned in the stimulus material section). The stimuli in the second block were presented in a different randomized order. Each decision made on the continuation of a video has a binary rating, as only the best decision was counted as 1 (correct) while all other options were rated as 0 (incorrect). At first, 5 different sample screenshots (example view see Fig A.1 in equirectangular form or S1

Figure A.1.: Example stimulus in equirectangular format.

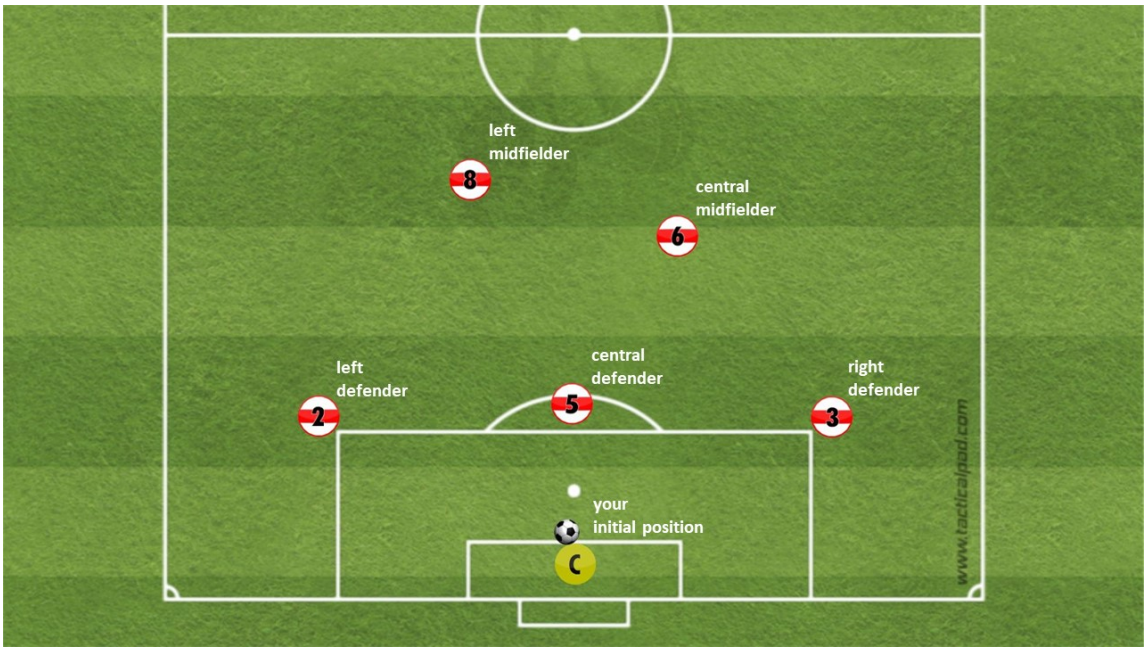

Figure A.2.: Schematic overview of the response options. Emergency option kick out is not shown.

Video for a cross-section of the stimulus presentation sphere) and the corresponding sample stimuli were shown and explained to acclimate the participant to the setup. To learn the decision options, we also showed a schematic overview before every sample screenshot (see Fig. A.2).

Eye Tracking

The raw data of the SMI Eye tracker can be exported from the proprietary BeGaze software as CSV files. BeGaze already provides the calculation of different eye movement features based on the raw gaze points. As we get high-speed data from the eye tracker, we use the built-in high-speed event detection. The software first calculates the saccades based on the peak threshold, which means the minimum saccade duration (in ms) varies and is set dependent on the peak threshold default value of $40°/s$. In a second step, the software calculates the fixations. Samples are considered to belong to a fixation when they are between a saccade or blink. With a minimum fixation duration of 50 ms, we reject all fixations below this threshold. As there is no generally applicable method for detection of smooth pursuits, this kind of event is included and encoded as fixations with longer duration and wider dispersion. We marked fixations with a fixation dispersion of more than 100 px as smooth pursuits. By doing this, we split fixations into normal length fixations and long fixations which we consider to be and refer to as smooth pursuits. This threshold is an empirical value based on the sizes of the players as the main stimuli in the video. The following section describes the steps that are necessary to train a model based on these eye movement features.

Feature selection

As it is not clear which subset of eye movement features explains the difference in expertise completely, we followed a brute-force method, considering all possible measures issued by the eye-tracking device and subsequently evaluating their importance. For the classification of expertise level we focus on the following features:

- event duration and frequency (fixation/saccade)
- fixation dispersion (in °)
- smooth pursuit duration (in ms)
- smooth pursuit dispersion (in °)
- saccade amplitude (in °)

- average saccade acceleration (in $°/s^2$)
- peak saccade acceleration (in $°/s^2$)
- average saccade deceleration (in $°/s^2$)
- peak saccade deceleration (in $°/s^2$)
- average saccade velocity (in $°/s$)
- peak saccade velocity (in $°/s$)

Each participant viewed 26 stimuli twice, resulting in 52 trials per subject. First, we viewed the samples of these 52 trials and checked the confidence measures of the eye-tracking device. We removed all trials with less than 75% tracking ratio, as gaze data below this threshold is not reliable. Due to errors in the eye-tracking device, not all participant data is available for every trial. Table A.2 shows an overview of the lost trials. For two participants, 11 trials had a lower tracking ratio; on participant 18, we lost 35 trials; and on participant 33, one trial was lost. This results in 1658 out of 1716 valid trials in total. 3.3% of the trials were lost due to eye-tracking device errors.

Overview erroneous trials

Participant	Number of valid trials
1	11
8	11
18	25
33	1
all others	0

Table A.2.: Overview of the amount of erroneous trials, based on eye-tracking device errors.

Data cleaning

We checked the remaining data for the quality of saccades. This data preparation is necessary to remove erroneous and low-quality data that comes from poor detection on behalf of the eye-tracking device and does not reflect the correct gaze. Therefore, we investigated invalid samples and removed (1) all saccades with invalid starting position values, (2) all saccades with invalid intra-saccade samples, and (3) all saccades with invalid velocity, acceleration, or deceleration values.

1. Invalid starting position: 0.22% saccades started at coordinates (0;0). This is an encoding for an error of the eye-tracking device. As amplitude, acceleration, deceleration, and velocity are calculated based on the distance from the start- to the endpoint, these calculations result in physiological impossible values, e.g., over 360°saccade amplitudes.

2. Invalid intra-saccade values: Another error of the eye-tracking device stems from the way the saccade amplitude is calculated through the average velocity (Eq A.1) which is based on the distance of the mean of start and endpoints on a sample-to-sample basis (see Eq A.2). 3.6% of the saccades had at least one invalid gaze sample and were removed (example see Fig A.3).

$$\oslash Velocity * EventDuration \tag{A.1}$$

$$\frac{1}{n} * \sum_{1}^{n} \frac{dist(startpoint(i), endpoint(i))}{EventDuration(i)} \tag{A.2}$$

On Fig A.3, the gaze signal samples 7, 8, 14-16, 18-20 (x-axis) both, the x- and y-signal (blue and red line, respectively) show zero values and thereby indicate a tracking loss. As the saccade amplitude is based on the average velocity which is calculated on a sample-to-sample Eq A.2, the velocity from samples 6 to 7, 8 to 9, 13 to 14, 16 to 17, 17 to 18, and 20 to 21 significantly increase the average velocity as the distances are high (on average over 2400 px for x-signal and over 1000px for y-signal, which corresponds to a turn of 225° on the x-axis and 187.5° on the y-axis in the time of 4 ms between two consecutive samples).

There are two interpretations for saccadic amplitude. The first refers to the shortest distance from start to the endpoint of a saccadic movement (i.e., a straight line) and the second describes the total distance traveled along

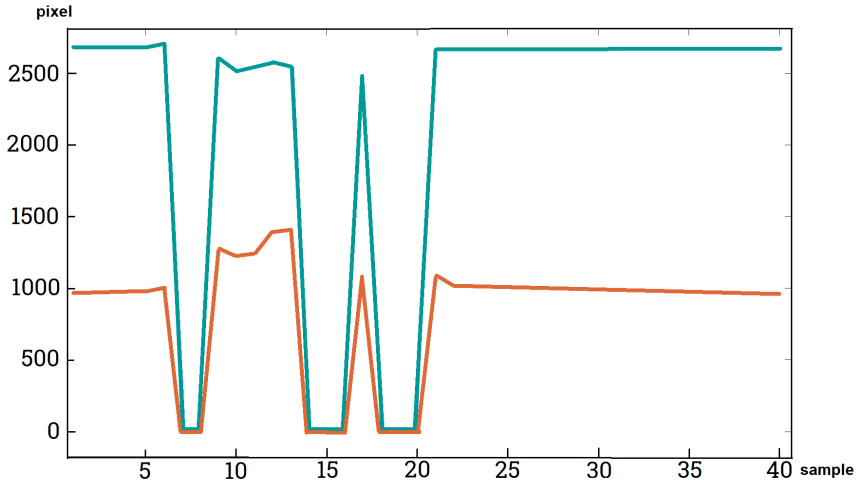

Figure A.3.: Example of invalid intra-saccade values. The x-axis shows the number of the gaze signal sample (40 samples, 250 Hz, 160 ms duration) and the y-axis shows the position in pixel. The blue line represents the x-signal of the gaze and the orange line the y-signal.

the (potentially curved [33], p.311) trajectory of the saccade. The SMI implementation follows the second definition. We could have potentially interpolated invalid intra-saccade samples instead of completely removing the complete saccade from analysis, however, this leads to uncertainties that can affect the amplitude depending on the number of invalid samples and does not necessarily represent the true curvature of the saccade.

3. As the velocity increases as a function of the saccade amplitude [141], 4.8% of the saccades were ignored because of the restriction on velocities greater than 1000°/s. Similar to extreme velocities, we removed all saccade samples that exceeded the maximum theoretical acceleration and deceleration thresholds. Saccades with longer amplitudes have higher velocity, acceleration, and deceleration, but can not exceed the physiological boundaries of 100.000 $°/s^2$ [33]. 3.0% and 4.0%, respectively, of all saccades that exceeded this limit. As most of the invalid samples had more than one error source, we only removed 5.5 % of the saccades (3.5% of all samples) in total.

After cleaning the data we use the remaining samples to calculate the average, maximum, minimum, and standard deviation of the features. This results in 36 individual features. We use those for classifying expertise in the following.

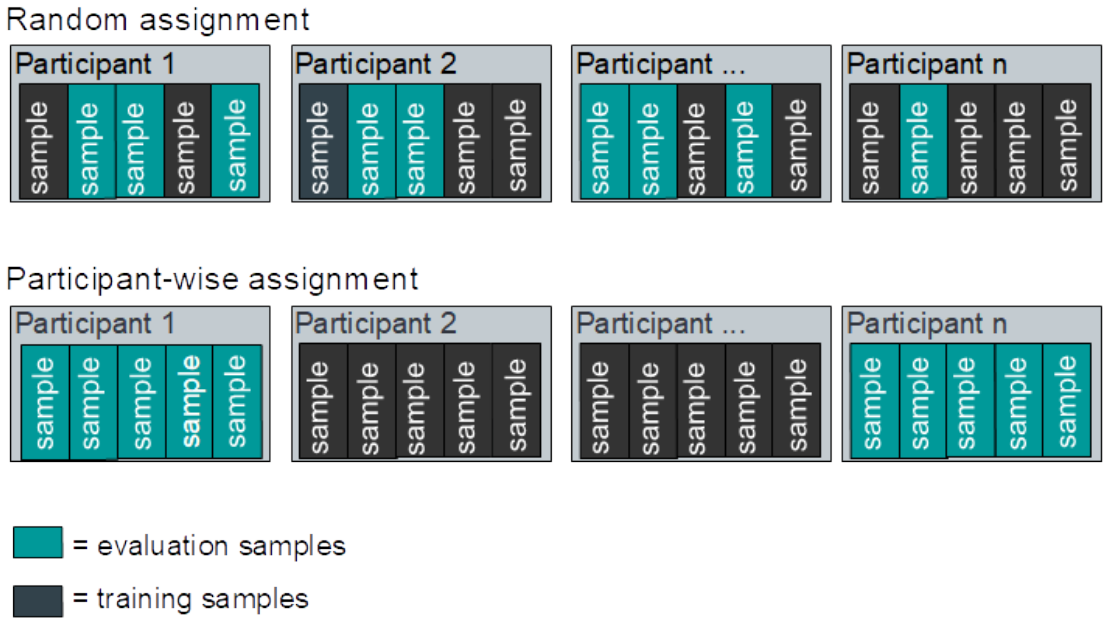

Figure A.4.: Example sample assignment. Top row shows a random assignment of samples, independent of the corresponding participant. Bottom row shows participant-wise sample assignment to training and evaluation set.

Machine learning model

In the following, we refer to *expert samples* as trials completed by an elite youth player of a DFB goalkeeper camp, *intermediate samples* as those of regional league players, and *novice samples* as those of amateur players. We built a support vector machine model (SVM) and validated our model in two steps: cross-validation and leave-out validation. We trained and evaluated our model in 150 runs with both validations. For each run, we trained a model (and validated with cross-validation) with samples of 8 experts, 8 intermediates, and 8 novices samples, and used the samples of two participants from each group of the remaining participants to predict their classes (leave-out validation). The experts, as well as the intermediates and the novice samples in the validation set, were picked randomly for each run.

Sample assignment

We found that the way in which the data set samples are split into training and evaluation sets is very important and a participant-wise manner should be applied.

By randomly picking samples independent of the corresponding participant, participant samples usually end up being distributed on the training and the evaluation set (illustrated in Fig A.4). This leads to an unexpected learning behavior that does not necessarily classify expertise directly, but, rather, matches the origin of a sample to a specific participant thereby indirectly identifying that participant's level of expertise. This means that a model would work perfectly for known participants, but is unlikely to work for unseen data. Multiple studies show that human gaze behavior follows idiosyncratic patterns. Holmqvist et al. [33] show that a significant number of eye-tracking measures underlay the participants' idiosyncrasy, which also means that the inter-participant differences are much higher than intra-participant differences. A classifier learns a biometric, person-specific measure instead of an expertise representation.

Model building

To find a model robust to high data variations, we applied cross-validation during training. The final model is based on the average of k=50 models, with k = number of folds in the cross-validation. For each model m_i, with $i \in \{1, \ldots, k\}$, we use all out-of fold data of the i-th fold to train and evaluate m_i with the in-fold data of the i-th fold (this procedure is illustrated in Fig A.5). The final model is evaluated with a leave-out validation. The cross-validation step during training is independent of the leave-out validation with totally new data (never seen by the model). Information from cross-validation is used during the building and optimizing of the model and leave-out validation solely provides information about the prediction accuracy of the model when using completely new data.

With a total of 810 valid samples, equally distributed on expert, intermediate, and novice samples, we built a subset of 552 samples for training the model and a subset of 258 samples for evaluation. As each sample represents one trial, our approach here is to predict whether a trial belongs to an expert, intermediate, or novice class. We tested assumptions in different approaches.

Classifiability

Firstly, we used all 46 features to check the classifiability of this kind of data. The first approach contains all features from section *Feature selection* A.1.2 with their derivations, (namely average, maximum, minimum, and standard deviation) to build an SVM model (Tables A.3, A.4 and A.5 show all features with their derivations, split by class). When the binary case (expert vs. intermediates) results point out classifiability, the ternary case (expert vs. intermediate vs. novice) should be investigated.

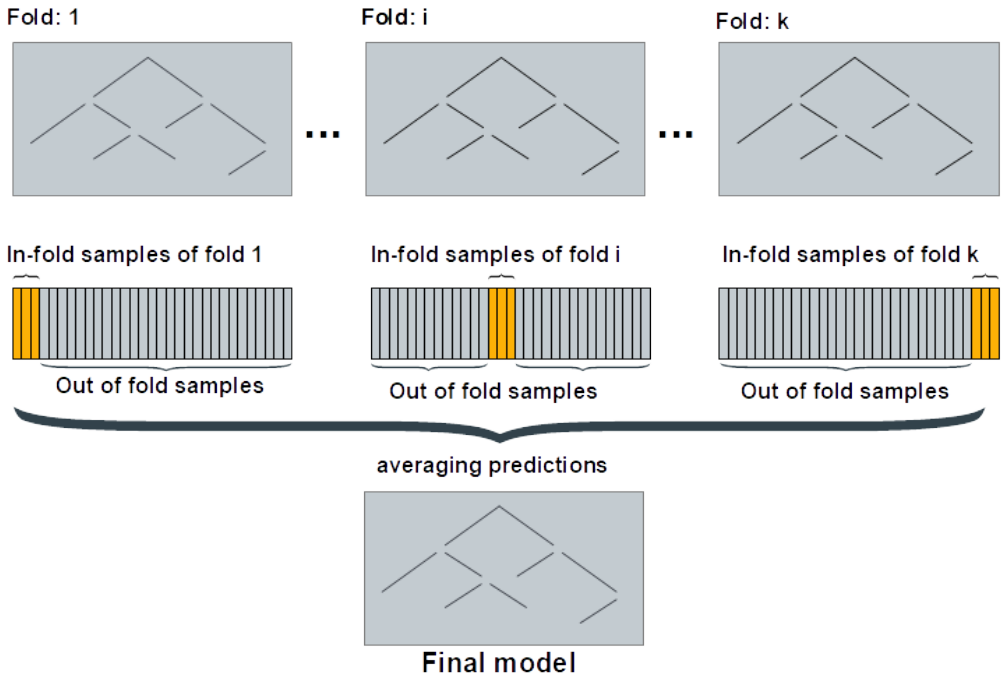

Figure A.5.: Illustration of the k cross-validation procedure. Each of the k models has a different out-of-fold and in-fold data set. We build the final model on the average of all predictions from all k models.

Significant features

Secondly, we had a look at the features themselves and checked for differences between the single features according to their class and as well as checking for the significance level of feature differences under 0.11%. We built a model based on the features that have a significance level under 0.11% (Tables A.3, A.4 and A.5 all white cells, gray cells mean there is no significant difference between the groups).

Most frequent features

In a third approach, we reduced the number of features by running the prediction on all 46 features 150 times. By taking the most frequent features in the model, we search for a subset of features that prevent the model from overfitting and allow for interpretable results representing the differences between expertise classes with a minimum amount of features. These most frequent features are imperative for the model to distinguish the classes. During training, the model indicates which features are the most important for prediction in each run. The resulting features with the highest frequency (and therefore highest importance for the model) in

Novices

	average	std. dev.	minimum	maximum
Fixation				
frequency (Hz)	0.21	-	-	-
duration (ms)	214.01	31.92	190.49	239.30
dispersion (pixels)	72.09	25.68	24.67	110.52
Saccade				
frequency (Hz)	0.07	-	-	-
duration (ms)	71.68	38.86	26.514	175.46
amplitude (°)	9.29	9.41	0.57	51.40
Saccade mean acceleration				
mean ($°/s^2$)	4263.38	2482.01	366.66	13984.56
peak ($°/s^2$)	9322.48	5777.27	231.83	28355.22
Saccade deceleration				
peak ($/s^2$)	-6848.10	4166.26	-35563.64	-411.76
Saccade velocity				
mean ($°/s$)	105.46	65.02	20.28	298.13
peak ($°/s$)	215.24	129.29	40.31	766.15
Smooth pursuit				
duration (ms)	302.63	278.11	75.62	1026.32
dispersion (pixels)	622.80	201.26	185.43	1085.90

Table A.3.: All 46 features with their derivations. Novice class. Green cells show features with significant differences between classes. Orange cells stand for the most frequent feature.

our test can be seen in Tables A.3, A.4 and A.5, in orange.

Intermediates

	average	std. dev.	minimum	maximum
Fixation				
frequency (Hz)	0.25	-	-	-
duration (ms)	255.22	53.37	215.83	299.62
dispersion (pixels)	73.17	26.54	23.07	114.76
Saccade				
frequency (Hz)	0.08	-	-	-
duration (ms)	84.34	59.72	26.12	246.12
amplitude (°)	9.88	10.674	0.57	54.83
Saccade mean acceleration				
mean ($°/s^2$)	4123.97	2685.99	315.34	15472.88
peak ($°/s^2$)	8920.17	5989.25	216.72	28266.00
Saccade deceleration				
peak ($°/s^2$)	-6948.49	4770.06	-36334.13	-231.35
Saccade velocity				
mean ($°/s$)	104.19	66.68	21.52	331.11
peak ($°/s$)	213.83	136.52	40.10	764.02
Smooth pursuit				
duration (ms)	291.09	278.71	73.83	977.12
dispersion (pixels)	425.08	124.85	168.32	694.37

Table A.4.: All 46 features with their derivations. Intermediate class. We consider samples as belonging to a smooth pursuit when the dispersion of the samples is greater than 100 px. As the size of the players in the stimulus varies around 90 pixel + a buffer.

Experts

Features	average	std. dev.	minimum	maximum
Fixation				
frequency (Hz)	0.24	-	-	-
duration (ms)	241.50	58.62	198.13	291.72
dispersion (pixels)	72.83	25.989	21.73	114.54
Saccade				
frequency (Hz)	0.00	-	-	-
duration (ms)	65.47	35.54	25.01	163.41
amplitude (°)	8.93	9.430	0.56	52.02
Saccade mean acceleration				
mean ($°/s^2$)	4769.65	3064.34	390.09	18965.94
peak ($°/s^2$)	10026.45	7094.930	175.24	39445.12
Saccade deceleration				
peak ($°/s^2$)	-7912.19	5492.28	-43479.91	-362.39
Saccade velocity				
mean ($°/s$)	110.67	72.73	21.18	375.36
peak ($°/s$)	238.37	157.74	40.26	935.51
Smooth pursuit				
duration (ms)	276.78	265.67	74.40	953.66
dispersion (pixels)	399.93	112.41	336.01	505.03

Table A.5.: All 46 features with their derivations. Expert class.

A.1.3. Results

We first report the results of an intra-expert classification test to see whether inter-experts differences are smaller than inter-class differences. Then, since we first

need to know whether there are differences between experts and novices, the classifiablity test (binary classification) provides a deeper analysis of the model trained with all features for distinguishing experts and novices. The remaining chapter describes two ternary models which are based on a subset of features obtained through 1) their significance level and 2) their frequency in the all feature model.

Intra-expert classification

To strengthen the implicit assumption of this paper that it is possible to distinguish between novices, intermediates, and experts based on their gaze behavior, we evaluated our expert data separately by flipping a subset of experts with intermediates. After 100 iterations in which half of the experts were randomly labeled as intermediates, the average classification accuracy was below chance-level, meaning the model can not differentiate between experts and flipped experts properly. This strengthens our assumption that inter-expert differences are smaller than inter-group differences between experts, intermediates, and novices.

Binary classification

The classifiability test shows promising results. This binary model is able to distinguish between experts and intermediates with an accuracy of 88.1%. The model has a false negative rate of 1.6% and a false positive rate of 18.6%. This means the binary model predicted two out of 260 samples falsely as class zero and 29 samples that are class zero as class one. As the false-negative rate is pretty low, the resulting miss rate is low (11.9%) as well. The confusion matrix (Fig. A.6) shows the overall metrics. The binary model is better in predicting class zero samples (intermediates) than class one samples (experts). The overall accuracy of 88.1% is sufficient to investigate a ternary classification. In the following, we show deeper insights on the ternary approaches by looking at accuracy, miss rate, and recall of the ternary models and compare those values between the All-feature model (ALL), most frequent features model (MFF), and significant features model (SF). This is to see if there is a better-performing model with fewer features.

Accuracy

The differences in accuracy between the three approaches are barely visible when looking at the median (ALL: 75.08%, MFF: 78.20%, SF: 73.95%), but even greater when comparing the 75th percentile (ALL: 80.989%, MFF: 85.44%, SF: 79.25%, see Fig. A.7). All models show a wider range of accuracy values which means these models might over fit more on some runs and under fit on others. The lower adjacent of all models is higher than the chance level (ALL: 53.46%, MFF:

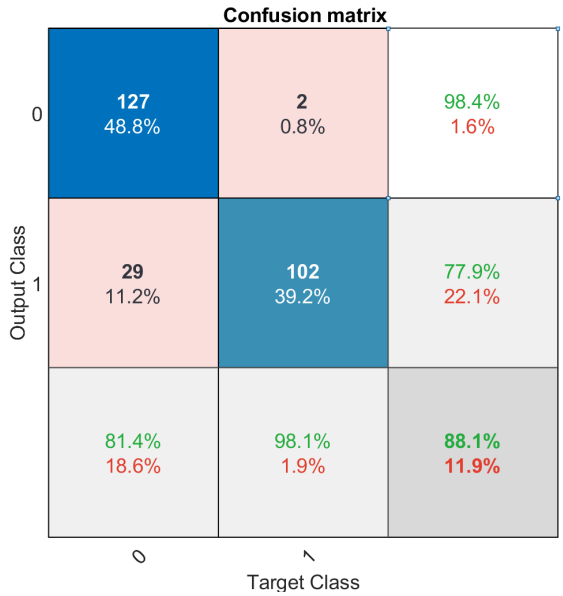

Figure A.6.: Binary confusion matrix about predictions on 100 randomized runs.

52.93% and SF: 52.41%), which means all models perform better as guessing. The chance level for 3 classes is 33.33%. A system that would only guess the correct class would usually end up with an accuracy of about 33.33%. Although not in each run, on average all models show a much better performance. Even the worst classification is over 20% higher than the chance level. Successful performance for classification expertise in machine learning models is usually when their average accuracy is between 70% and 80%. A statement about the performance of a model with lower than 70% accuracy depends on the task and how much data is available. Sometimes there are only a few people in the world who can be considered experts. As the accuracy is a rough performance metric that only provides information about the number of correct predictions (true positives and true negatives), we offer a more detailed look into the performance of the methods by comparing the miss rates of the single approaches.

Miss rate

The miss rate is a metric that measures the rate of wrongly classified samples belonging to class x but predicted to belong to class y. The models are better at predicting the membership of samples belonging to expert and intermediate classes than the novice class. This results in miss rates that are only a little lower than the chance level when looking at the median miss rates (All: 28.12%, MFF: 23.81% and SF: 26.80%, see Fig. A.8). The upper adjacent shows a high range of miss

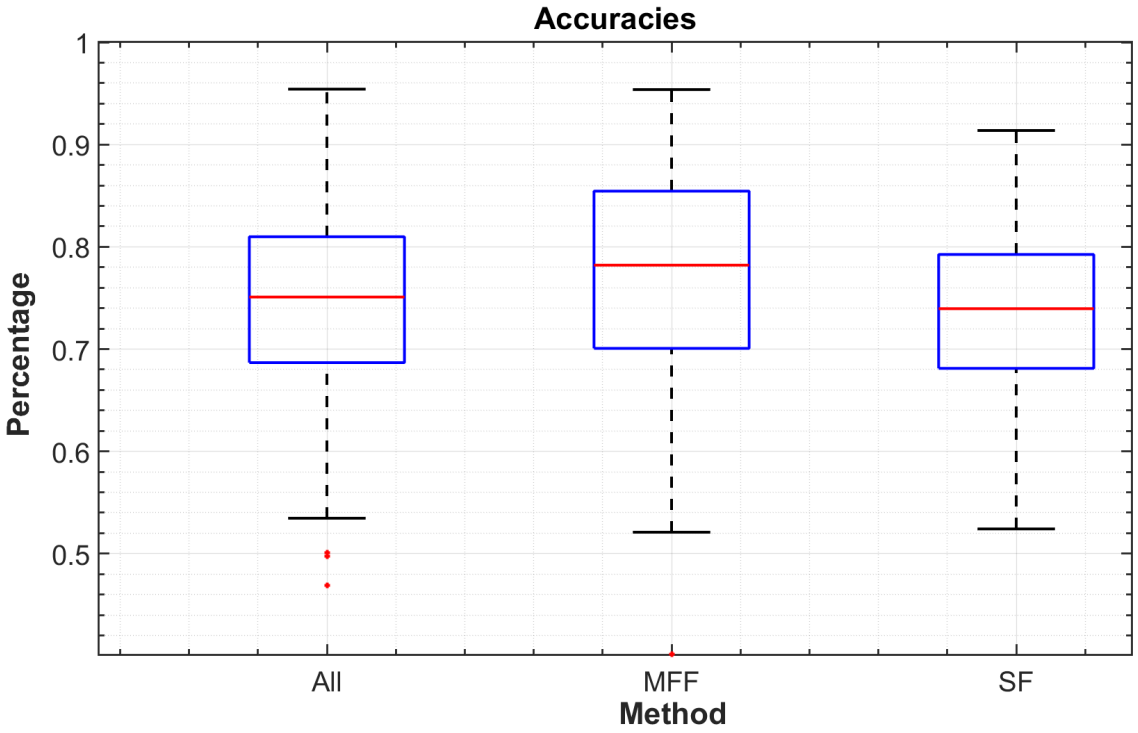

Figure A.7.: Box plot showing the accuracy values of the ternary methods. All three models have median accuracy values $\sim 75 - 80\%$

rates reaching even values of over 43.19% for the SF-model. The MFF-model has the lowest median miss rate of all three methods with a miss rate of 41.96%.

Recall

Recall provides information about the rate of predicted samples belonging to class x in relation to the number of samples that really belong to class x. All three models have a median recall of over 70% (as can be seen in Fig. A.9). In the ternary case, the chance level is at 33.33% which means all models have a recall over two times higher than the chance level as the lower adjacent of all three models is higher than 33.33%. The MFF-model median is the highest at 76.18% followed by the SF-model at 73.19% and the ALL-model at 71.87%. Again the MFF-model has the best performance values of all three methods.

Most frequent features

The most frequent features in 100 runs are summarized in Table A.6. Only the minimum of the saccade duration has $p > 0.011$. This means the differences are

not statistically significant. All other features show significant differences, signifying that a Mann-Whitney-U-test discards the null hypothesis that there are no differences with $p < 0.011$ for each of the features.

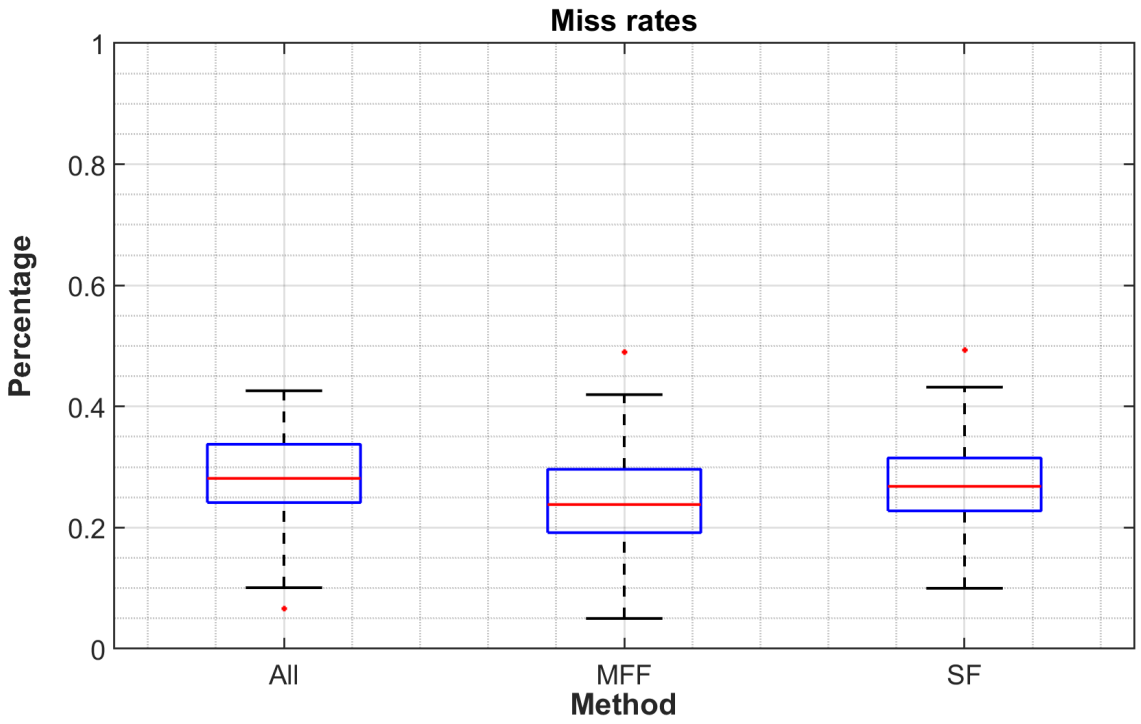

Figure A.8.: Miss rates of ternary methods.

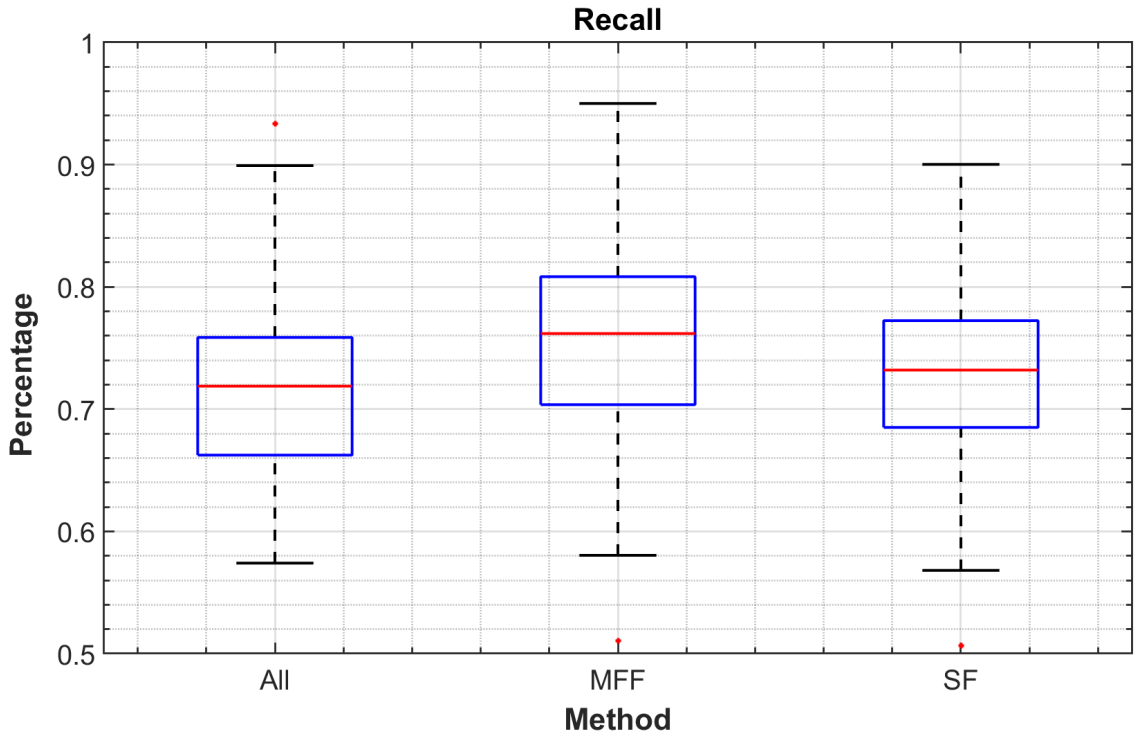

Figure A.9.: Recall values of ternary methods.

Most frequent features

	derivation	novice	intermediate	expert	p-value	significant?
saccade duration (ms)	std. dev.	38.869	59.726	35.548	3.33*e-08	1
saccade duration (ms)	minimum	26.514	26.127	25.019	0.242	0
peak saccade deceleration ($°/s^2$)	std. dev.	4166.262	4770.063	5492.287	2.49*e-18	1
peak saccade velocity ($°/s$)	std. dev.	129.294	136.529	157.740	6.19*e-07	1
smooth pursuit disp. (pixels)	average	622.805	425.089	399.939	9.66*e-82	1
smooth pursuit disp. (pixels)	minimum	185.437	168.320	336.016	5.44*e-12	1
smooth pursuit disp. (pixels)	maximum	1085.903	694.370	505.031	1.52*e-81	1

Table A.6.: **All most frequent features.**

A.1.4. Discussion

In this work, we have presented a diagnostic model to classify the eye movement features of soccer goalkeepers into expert, intermediate, and novice classes. We further investigated how well the features provided by the diagnostic model led to explainable behavior. Our model has shown that eye movement features are highly informative and well suited to distinguish different expertise classes. Based on a support vector machine as a simple machine learning model, we were able to classify three different expertise groups at an average accuracy of 78.2% (compared to the baseline of 33.3% in a three-class classification problem), thus a quality result for current machine learning techniques. As the performance values differ, the real-world application has to be further evaluated with larger subject groups. A closer look at the classification results reveals that our model can distinguish correctly between experts and intermediates. This is due to the fact that experts and intermediates have already been tested in the sense that they play in higher leagues and have already proven their ability. Thus, there is ground truth for these classes. A limitation of the classification model is currently the novice group. Since our novice group consists of participants with no regular training or involvement in competitions, novices can be equally talented players regarding their gaze behavior who have simply not yet proved their ability in a competition. This assumption is especially evident in the false-negative rate of 1.6% and the false positive rate of 18.6% from the binary model, respectively. This means that 18.6% of novice samples are classified as intermediate samples, but only 1.6% of the intermediate samples are classified as a novice. As is usual in expertise research, a proportion of low performers (novices) can also be found in higher classes. Our models confirm that the correct classification of novices is considerably more difficult than other classes since there is, to date, no objective ground truth. Despite this limitation, our model achieved a very good average accuracy of 78.2%. Most likely, a model with more subjects and finer graduation of the novices would offer a much better result. Machine learning models are data-driven and therefore can learn more from more data. However, the number of elite youth goalkeepers in Germany who can provide samples for the expert class is highly restricted. Out of 56 in total, we collected data from 12 for our study. An additional step would be to define a more robust ground truth for participants classified as novices. As it is more important that the model does not downgrade participants with higher expertise to a lower class, it can still be used as a diagnostic model. As aforementioned, the false positive rate only shows, that some novices with limited experience can perform better than others and therefore be classified into a higher class. This is correct because their gaze behavior is closer to intermediates than it is to typical novices.

By examining the individual eye movement features in more detail we have shown that, on the one hand, a subset of features is sufficient to create a solid clas-

sification and, on the other hand, that the differences in eye movement behavior between the individual groups are difficult to interpret. We only investigated the most frequent features since these features built the best-performing model. The differences are noticeable, but hard to interpret as there is no simple characteristic behind these features.

There are indications that 1) experts (std. dev. 35.54 ms) as well as novices (std. dev. 38.86 ms) have a more homogeneous saccade behavior compared to intermediates (std. dev. 59.72 ms). The lengths of the saccades differ less. However, it would be a fallacy to attribute the same viewing behavior to novices and experts due to the similar standard deviation and minimum duration of the saccades (novice: 26 ms, intermediate: 25 ms, expert: 25 ms). It is clear that both groups have similarly long saccades, but the novices have similarly long saccades and the experts similarly short saccades. Conversely, this means that the experts might have longer fixations than the novices and intermediates. These findings are in line with Mann et al. [20] who show that experts are over-represented in fewer, but longer fixations. Their visual strategy is often based on longer fixations to avoid saccadic suppression (which might lead to information loss). In our statistics, fixation durations did not exhibit to have significant differences between the three groups. This is in line with the findings of Klostermann et al. [42]. It also might be based on the split of the fixation values in short fixations and smooth pursuits. The source of these differences may also be the age difference between the single groups (see Table A.1). With the current data, this is not rigorously answerable.

Further differences between the groups can be found in the maximum peak deceleration of the saccades. There is a continuous increase in the maximum deceleration speed of the novices' saccades (4166.262 $^\circ/s^2$) to intermediates(4770.063 $^\circ/s^2$) to experts (5492.287$^\circ/s^2$), which is in line with the findings of Zwierko et al. [109] who found that the deceleration behavior can be inferred from different expertise classes.

One observation made by the experimenter during the study was that novices often follow the ball with their gaze for a long time. This behavior is less evident among experts. They tend to only look at the ball when it has just been passed or when they themselves are not in play. At these times, the ball can not change its path. This observation is supported by the values of the smooth pursuit dispersion. With 505.031 pixels maximum and 336 pixels minimum, experts have a very narrow window of smooth pursuit lengths. Basically, the maximum smooth pursuit of the experts (505.03 pixels) is less than half as long as the novices (1085.90 pixels), and the minimum smooth pursuits (expert: 399 pixels, intermediate 425 pixels, novices 622 pixels) is still 1/3 shorter than the novices. The intermediates are placed in the middle between the two groups. Again, the values are continuously decreasing. Based on the continuity of the average smooth pursuits that correlate negatively with the classes, as well as the maximum and standard de-

viation, it can be concluded that experts tend to make smooth pursuits of a more regular length. One explanation for this could be that, in addition to the opponents and players, the ball, as an almost continuously moving object, attracts a high level of attention. In order to maintain a clear overview in the decision-making process, soccer players are taught the following behavior: Shortly before the ball arrives at the pass goal, you look at it. This is done until the ball is passed away. Since the path of the ball can only be changed by a player who is in possession of the ball and not in the middle of a pass, it is only necessary to follow the path of the ball at the beginning and end of the pass. In the meantime, players should scan the environment for changes to keep track of options in the field. This leads to short smooth pursuits around the ball before the end and at the beginning of each pass so that experts can appreciate the ball and follow the ball with similarly long smooth pursuits. On the other hand, as aforementioned before, novices often follow the ball's path almost continuously or, at least, very often. The characteristics of the smooth pursuit support this theory. The characteristics of smooth pursuits differ significantly from one another in the three groups with an average, minimum, and maximum significant p-value of less than $1 * 10^{-}12$. The novices with 622.81 pixels make, on average, much longer smooth pursuits than the intermediates (525.09 pixels) and significantly more than the experts (399.93 pixels). With 185.44 pixels, the shortest smooth pursuits of the novices are smaller than those of the intermediates (168.32 pixels) and the experts with 336.01 pixels. The maximum values show a uniform behavior. With 1085.9 pixels, the novices have the highest maximum values after the intermediates with 694.37 pixels and the experts with 505.03 pixels.

Although the standard deviation of the lengths of the smooth pursuits does not belong to the MF features, clear differences can be seen here as well. The dispersions of the smooth pursuits with 201.27 pixels scatter far more among the novices than among the intermediates (124.85 pixels) and experts (112.41 pixels). These findings lead us to believe that a stimuli oriented investigation on gaze distribution for expertise recognition might reveal even more pronounced differences, i.e correlation between ball movement and smooth pursuits.

A.1.5. Conclusion & Implications

After the ternary classification of expertise, the next step should be the evaluation of a more robust classification model. As machine learning techniques are data-driven, adding more subjects to each group should, presumably, provide better results. As soon as a robust model is built, a finer-grained gradation should be considered to achieve a more sensible model that allows for the classification of participants in more classes by predicting their class in a more nuanced fashion. In our further work, we plan to expand our data set to more subjects in the current

groups, add more nuanced classes and add a physical response mode to infer speed and correctness in a standardized, controllable and objective manner, thus increasing the immersion. however, a fully interactive mode will only be possible when CGI can provide high enough quality and cost-efficient environments. Another step is to focus on the research of person-specific, gaze-based expertise weakness detection. As soon as a robust model is achieved, another point is to integrate the model into an online diagnostic system. To use the model online, the gaze signal can be directly drawn online at 250 Hz from the eye tracker by using the provided API of the vendor. Using a multi-threaded system, the data preparation and feature calculation can be done directly online in parallel to data collection. Only the higher level features (e.g. std. deviations) need to be computed when the trial ends and fed as a feature vector to the already trained model in order to estimate the class of the current trial. As predicting is completed by solving a function, the prediction result is supposed to be available a few moments after the trial ends. This is necessary as the prediction is the input for the adaption of the training. This work will be implemented in an online system for real-time gaze-based expertise detection in virtual reality systems with an automatic input for the presentation device to ensure dynamic manipulation of a scene's difficulty. With a prototype running in VR, we are planning to expand the system to be used in-situ with augmented reality glasses (AR). This may further pronounce the differences and lead to even better classifications. A more sensible model would allow, by mapping expertise on a larger number of classes, the dynamic manipulation of the difficulty level of a training system exercise or game level in virtual environments. Next to a training system for athletes and other professional groups, the difficulty level in a VR game can be dynamically adjusted based on the gaze behavior of the user. We are, however, aware that the small sample size restricts potential conclusions that can be drawn and may lead to contentious results. Another limitation of this work is the restriction presented by head movement unrelated eye movement features and the absence of a detailed smooth pursuit detection algorithm, which might be important. Therefore, in our future work, we will implement an appropriate event calculation method i.e. based on the work of Agtzidis et al. [142]. This work, however, strengthens the assumption that there are differences between the gaze behavior of experts, intermediates, and novices, and that these differences can be obtained through the methods discussed.Using machine learning techniques on eye-tracking data captured in a photo-realistic environment on virtual reality glasses can be the first step towards a virtual reality training system (VRTS). Objective expertise identification and classification leads to adaptive and personalized designs of such systems as it allows for a definition of certain states in a training system. A VRTS that can be used at home and, based on its objective and algorithmic kind, allows for self-training at home. The choice of difficulty can be adapted based on the expertise of the user. For higher-skilled users, the level of difficulty can be raised by pointing out fewer cues or showing more crowded,

faster/more dynamic scenes to increase the pressure placed on decisions. With enough data, it is also possible to adapt the training level based on personal deficiencies discovered during expertise identification in a diagnostic system. This can result in a system that knows a user's personal and perceptual weak spots to provide personalized cognitive training (e.g. different kinds of assistance like marking options, timing head movements, showing visual and auditory cues). Such a system is also potentially applicable in AR as the findings on the photo-realistic VR setup can be used in AR settings (i.e. in-situ). For uses such as AR-trainings - that can enhance physical training - the fundamental findings must be based on real gaze signals. As a second step, training systems can be developed based on the diagnostic findings. As, in addition to physical training, perceptual-cognitive training forms are increasingly being researched [143]–[146].

A.2. Differentiating Surgeons' Expertise Solely by Eye Movement Features

Abstract

Medical schools are increasingly seeking to use objective measures to assess surgical skills. This extends even to perceptual skills, which are particularly important in minimally invasive surgery. Eye tracking provides a promising approach to obtaining such objective metrics of visual perception. In this work, we report on results of a cadaveric study of visual perception during shoulder arthroscopy. We present a model for classifying surgeons into three levels of expertise using only eye movements. The model achieves a classification accuracy of 84.44% using only a small set of selected features. We also examine and characterize the changes in visual perception metrics between the different levels of expertise, forming a basis for development of a system for objective assessment.

A.2.1. Introduction

Arthroscopy is a popular minimally invasive surgical procedure that improves patient outcomes while at the same time conserving hospital resources. According to Monson et al. [147], patients experience less pain, have fewer complications and recover faster than with traditional open surgery. However, a surgeon needs advanced technical skills for this type of operation [148]. Arthroscopy involves inserting instruments and a scope into the joint (e.g. shoulder or knee) through small incisions. A key capability in performing arthroscopic surgery is the ability to use the scope to navigate through complex anatomy of the joint for inspection, diagnosis, and to locate the surgical site. The scope can rotate in multiple dimensions and casts its image on a screen placed next to the patient, which surgeons largely rely upon during surgery. Navigation is challenging due to complex anatomy, limited field of view, projection of the 3D space onto the 2D monitor, and the rotation of the monitor from the instrument plane.

Due to these technical challenges, there is growing interest within the medical community to optimize training, including having objective measures of performance for tasks like navigation. Since navigation is a psychomotor task in which visual perception plays a crucial role, it is natural to look to eye tracking for such a measure. Indeed, the role of eye movements is increasingly being investigated in surgery [148]. In particular, the role of eye movements is increasingly being investigated (for an overview see [148]). To determine whether eye tracking can serve as a basis for an objective measure in arthroscopy, first it must be determined whether, and to what extent, differences in surgeons' expertise are reflected by their eye movements. The findings from this study are significant for the design of

adequate training and evaluation scenarios for perceptual-cognitive diagnostic and training systems. In this work, we consider the perception of surgeons using eye movement patterns from three expertise levels in a human cadaveric study of diagnostic arthroscopy of the shoulder. We selected this task since it focuses on navigation skill in which perception plays a major role. We use stimulus-independent eye movement patterns to develop a model to classify the subjects into the three levels of expertise. Using only a small number of selected features, our model achieves a classification accuracy of over 84%. We further investigate differences in eye movement patterns among the three classes in order to understand how these patterns evolve with increasing levels of expertise. We hope that such an understanding can assist in developing specialized training to provide the appropriate support to surgeons at different expertise levels.

A.2.2. Related Work

In eye tracking studies, using artificial forms of presentation like virtual reality (VR) [149], [150] or images [99], [151] could omit important perceptual details requiring the participants to fill in through inference which often subsequently leads to the higher levels of frustration [150]. To provide a presentation mode that is as natural as possible, we use so called soft cadavers that provide natural tactile sensation while maintaining the naturalness of the scene. Although remote eye trackers are commonly used in lab studies [149], as soon as the participant changes to another direction (e.g. down at the cadaver), they can no longer capture the gaze signal. To allow the participant to use normal gaze behavior and move freely without data lost, we use a head-mounted eye tracker in combination with a 4k-screen. This setup supports natural gaze behavior as well as high control of the stimulus allowing us to capture highly detailed information of the tissue on a screen with high resolution and gaze signals on the cadaver, both with the same field camera. Eye tracking studies in surgery are differed in how they evaluated the gaze signal. The gaze signal on the stimulus was considered, i.e. target gaze behavior, switching behavior (alternating gaze between target and instrument), or following behavior (eye following the instrument) [149]. Other studies focused on quiet eye periods [71]. However, there are also studies that have gained insights at the feature level. For example, Kocak et al.[112] used stimulus-independent eye features in their binary classification and found significantly lower saccade rates, as well as significantly higher peak velocities for experts, which was confirmed by other studies [148]. Tien et al. [152] found a higher fixation rate in experts. Eivazi et al. [99] show differences in time to first fixation and mean fixation duration. However, theses differences were not confirmed by Sondergren et al. [151], as in both studies fixation durations are analyzed differently and the choice of regions of interest plays an important role. These results show that eye movements can be

used to assess the surgical expertise and to define differences between groups.

Many studies have focused on the detection of differences in expertise between experts and novices [71], [150] and only few studies have focused on the development of eye movements. Studies focusing on development have used mostly simulators [112] or images [151]. Hidden Markov models (HMM) used in the latter study reveal differences in eye movement patterns between high and low performers. So far, several algorithms have been introduced to eye tracking including supervised methods like support vector machines [29], [100] and neural networks [30]. Ahmidi et al. [153] mixed instrument movements with eye movement data and achieved a binary classification accuracy of 82.5% for skill level classification. All these studies show that eye movement data can be used to differentiate between experts and novices and that it is not necessary to determine exactly where the surgeons were looking to measure their skill accurately.

A.2.3. Participants and Methods

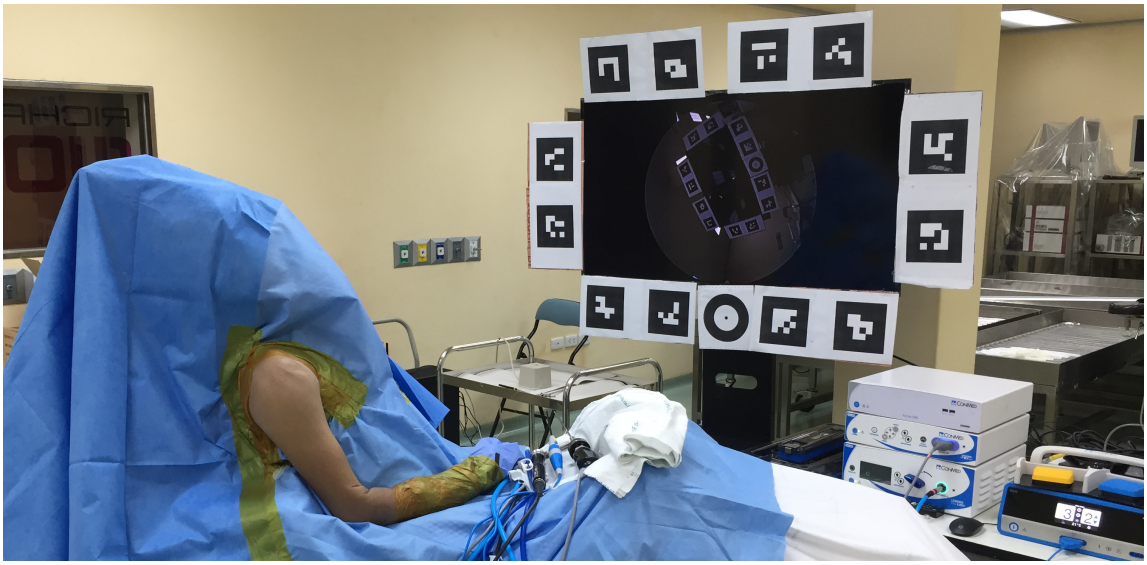

Figure A.10.: Experimental setup showing cadaver, arthroscopic equipment, and 4k monitor with ARUCO markers.

Procedure

This work makes use of the eye tracking data set from the work of Yin et al. [110]. Their data set contains eye movement data for three classes of surgeons: 3rd year residents (R3), 4th year residents (R4), and fellows. Each class consists of five (n=5) participants, equally. We even use the data of the two participants that were

left out in their study because of a gaze signal offset. Since we only use relative features, we can use the data of these participants too. In their study, participants were placed in front of the cadaver and four feet away from the 4k, 52-inch screen where the output of the arthroscope was displayed (Figure A.10). Each participant was familiarized with the setup and asked to navigate and diagnose 12 anatomical landmarks in the shoulder, while wearing a Tobii Pro Glasses 2 eye tacker. The gaze was recorded with Tobii software.

Data preparation

The Tobii Glasses 2 were set to a frame rate of 100 Hz, thus a gaze sample is available every 10 ms and saved with a timestamp, x-, and y-coordinates. The samples are used to calculate fixations and saccades metrics using the Tobii Fixation Filter, with a sliding window averaging method and the feature classification algorithm. These samples are used to calculate metrics like fixations and saccades. To calculate these metrics, we used the Tobii Fixation Filter, using a sliding window averaging method. The feature calculation is based on the classification algorithm of Olson [154] with a default velocity threshold of 0.7 pixels/ms. The raw eye tracking data, as well as the fixations and saccade metrics, are exported from the Tobii Studio software. From the fixations, velocity of the saccades, saccade duration, values of the gyroscope (yaw, pitch and roll) as well as the amplitude of saccades, we use the person-specific average, minimum, maximum and standard deviation as features. While Tobii provides metrics about the first saccade and first fixation too, we did not include them. Since our participants were familiarized with the glasses for different lengths of time when the trials started, we end up with chaotic first saccades, which have no informative character.

As our aim is to infer which features contribute to expertise differences, we first used all the exported features from the Tobii Studio Software and added common metrics to them. Subsequently, we evaluated their frequency in the model building process and rated the most frequent used features to build a model with this subset of features for expertise acquisition. To incorporate uncertainties, we trained the model 150 times and calculated the most frequently used features by taking the features with the maximum number of occurrences in the training process. We added certain typical eye movement features which we calculated by ourselves. The fixations were split into small fixations and smooth pursuit fixations. As the Tobii Software does not provide calculations of smooth pursuits, which are assumed to help differentiating different expertise classes, the smooth pursuit events were encoded in the fixations. We therefore treated fixations with a dispersion over 30 pixels as smooth pursuit fixations. This threshold was empirically defined during data analysis.

The set of features was:

- Saccade duration (average, min, max. std. dev.)
- Fixation duration (average, min, max. std. dev.)
- Smooth pursuit dispersion (average, min, max, std. dev.)
- Fixation frequency
- Saccade frequency
- Pupil diameter (average, min, max. std. dev.)
- Gyroscope X,Y,Z (average, min, max. std. dev.)

We decided to include the gyroscope values because they could provide information about head movement between screen and cadaver may be revealed. The integration of pupil diameter features is based on the assumption that experts may have less fluctuating pupil diameter since their mental effort is considered to be smaller. Vice versa, the pupil diameter of intermediates and novices may reveal expertise differences by such effects.

Machine learning model

We used all 38 features to build a support vector machine (SVM) model in 150 independent runs. On each of the 150 runs we keep out one participant (leave-one-out validation). This participant is our test set and has never been seen by the model (of the current run) before. Therefore, in each run we take all data of the remaining 14 participants to train the model and test it with the unseen data of the test set participant. While the training algorithm iterates over the same procedure it changes the participant for the test set (sequentially iterating over the participant numbers from 1 to 15) 150 times. Thus, each participant is used as test set 10-times in total. By having 10 runs for each participant, we are taking statistical fluctuations into account. To ensure independence between runs, we train a new model on every run and report the accumulated accuracy values of the 150 runs. Thus, in each run, the model is trained with 14 participants and tested with the test set data of one participant, which is unseen by the current model. A strict separation of data in a participant-wise manner is very important, as mixing up samples of one person into training and testing data would allow the model to remember person-specific (idiosyncratic) features and restrict a real expertise learning process.

On each run, the data of the 14 participants of the training set is split into 10-folds. This is called a 10-fold cross-validation. The cross-validation is important

to protect the model against over fitting. In each fold, $\lceil \frac{1}{10} \rceil$ of the 14 participants that belong to training set is used to validate the model that is trained with $\lfloor \frac{9}{10} \rfloor$ of 14 participants. Which participant belongs to training or validation set, is decided randomly. However, the split is always done participant-wise to prevent an idiosyncratic learning behavior of the model.

In a first model, we use all 38 features to check the classifiability of the data set and afterwards reduce the amount by taking the four most frequent features of 150 runs. The most frequent features are features that have the highest importance values for a single model prediction. In each run we built a queue of all 38 features sorted by importance for the current model. Subsequently, we computed their overall frequency over all models.

A.2.4. Results

Our first classification model shows promising results with an average accuracy of 60%. As a system that would simply guess the class, would only reach a chance-level of 33.33%, the all feature model can already be considered as well-performing. But as we want to specify the results to allow a precise statement about a high performing classification with the least amount of features, we continued by collecting all features and their importance values on 150 runs of the all feature model and took the most frequent features (MFF) as a new set. With this subset of four features, shown in Table A.7, earlier counteracting features may be avoided and a precise statement about the differences of the groups can be stated. The final SVM model with the four MFF uses a linear kernel and a box constraint of 11.0174. We adopted one-vs-all approach for multi-class classification with the kernel scale remains 1. Before training, we standardized the data. Training took

Most frequent and important features

	Feature	derivation
1.	Peak velocity of saccades	standard deviation
2.	Amplitude of saccades	minimum
3.	Total amplitude of saccades	sum
4.	Saccade duration	standard deviation

Table A.7.: The most important and frequent features on 150 runs.

about 56.03 sec.

Performance metrics

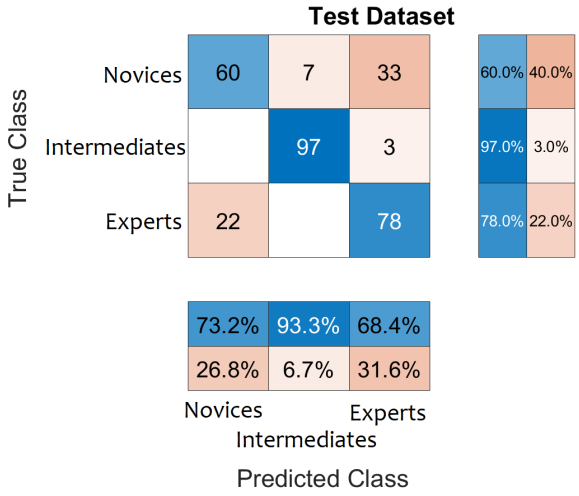

Figure A.11.: Performance values on 100 runs.

With an accuracy of 84.44% the model improved over 20 percentage points, compared to the all feature model. Figure A.11 shows the confusion matrix after 100 runs. 7 samples of the novice class were classified as intermediate and 33 as expert. This results in a class accuracy of 60%. The classification of the intermediates peaked at 97%, as only 3 samples were classified as experts and non as novices. This is especially interesting since the intermediates are in between the other classes and are therefore more likely to spread to both sides. The expert samples were with 78 samples correctly and 22 samples as novice samples, the second best classified class. The average recall is with 95.43% extremely high which is confirmed by the average miss rates. Only 4.57% of samples were misclassified. For the SVM model we achieve an area under the curve (AUC) of 0.91.

Feature evolution

As we consider there is a cognitive process going on, forming the optimal gaze behavior from novice to expert, we have a look at the evolution of the gaze features between the classes to describe such process as good as possible. To analyze these evolutionary steps, we have a look at the single feature characteristics separately. We do that with the four MFF from Table A.7. Table A.8 contains the characteristics of the four MFF features.

Average Feature Evolution

	Fellow	R4	R3
Saccade peak velocity (STD)	93.26 °/s	121.72 °/ s	117.45 °/s
Saccade amplitude (min)	0.86 °	0.40 °	0.64 °
Total saccade amplitude	481.32 °	1120.74 °	1956.21 °
Saccade duration (std. dev.)	18.96 ms °	16.58 ms °	23.54 ms °

Table A.8.: Average feature evolution between classes.

The table shows that experts have a smaller standard deviation of the peak velocity of the saccades (93.26 °/s). This feature is hard to interpret, but one assumption may be that experts have a more uniform distribution of saccade velocities. This means they do more saccades at the same speed, in a structured and planned way, compared to intermediates and novices. Interestingly, intermediates as the middle class between expert and novice show a much more diverse saccade peak velocity behavior (121.72 °/s). Novices are in the middle between experts and intermediates. A higher value for the standard deviation of the saccade peak velocities could be an indicator for a more chaotic gaze behavior, but it is hard to draw a conclusion about such a feature. When having a look at the minimum saccade amplitudes, we can see the same differences. The experts have on average a larger minimal saccade of length 0.86 °, compared to the intermediates with 0.40°and the novices with 0.64°. Again, we can see that the novices are in between the experts and intermediates. Only the total amplitude of all saccades shows a uniform evolution. The experts do a total of 481.32°of saccade length, where intermediates do more than twice the experts (1120.74°) and novices (1956.21°) even more than five time the experts and nearly double the intermediates. Another interesting feature evolution can be seen in the standard deviation of the saccade durations. This feature is also hard to interpret, but one possibility could be that experts with 18.96 ms and intermediates with 16.58 ms have slightly more order in their saccades than novices. Though the differences are very small and should be confirmed with more data.

A.2.5. Discussion

In this work we developed a model with supervised machine learning techniques that is able to distinguish three levels of expertise solely on the basis of eye movements during an arthroscopic surgery of the shoulder. With an accuracy of 82.33% the model can be considered as performing well on this 3-class problem. Thus, it can be stated that expertise differences between three different groups of expertise are reflected by their eye movements. To further understand the differences between the three levels of expertise, we had a look at the four most frequent features of the model and analyzed the evolution of the characteristics between the groups. Except for the total amount of saccade amplitudes, the remaining three of the four most frequent features show a uniform evolution. First, novices tend to have a more chaotic gaze behavior and distribute their gaze over a larger portion of the scene by making many different saccades with different speed. They also tend to look more at the outside than the center. The evolution to intermediates shows an atypical behavior, as they tend to still gaze over a larger area of the scene than the experts, but do smaller saccades with a still diverse velocity. This might indicate, that they try to focus on more specific visual clues and start to concentrate on the center of the scene. In the next evolution step, the saccade velocities shrink significantly, which signifies a more planned scanning behavior, with somewhat longer saccades, concentrated more on specific areas. To summarize our findings, one can state that the evolution of novices to intermediates first tends to lead to a partly more chaotic gaze behavior, then turning to be more precise. With the investigations on the evolutionary steps, we can also define class dependent weak-spots in perception for each class. An evolution between the single classes is clearly recognizable. Thus, opening the way to a class-specific training system that is optimized for different steps in perceptional evolution. We also showed that for a high accuracy classification there are not many features needed. A subset of four features describing the gaze behavior is already enough to distinguish different classes. Luckily, the four features are easily calculated, which would allow the usage of the classification as an online classification system. Though, the classification would need to be done segment-wise after a certain period of time.

Further steps are to add more participants to each class, and refine the number of classes. This would allow a much finer classification and therefore a better understanding of the differences between the levels of expertise. A finer classification is important to robust assumptions made by the model about gaze behavior and optimize the recognition of class-specific weak-spots to be used in a training system.

A.3. A Study of Expert/Novice Perception in Arthroscopic Shoulder Surgery

Abstract

Arthroscopic shoulder surgery is an advanced orthopedic surgical procedure, which is particularly challenging due to the complex anatomy of the shoulder, and tight spaces for navigation, which also limits the view from the arthroscope. In carrying out arthroscopy, the ability to quickly and effectively navigate through the joint to reach a desired location is essential. Novices often experience confusion in trying to triangulate the information from arthroscopy output with the background knowledge of anatomy while orienting and navigating the instruments. In this paper, we report on the results of the first cadaveric eye-tracking study of arthroscopic surgery in which we investigate differences in perception between experts and novices. Novices' perception is analyzed with cognitive load analysis throughout the procedure and specifically, during the portions of the procedure in which subjects are observed to be confused. In investigating such portions, the gaze data analysis is supplemented with head rotations and acceleration information from gyroscope and accelerometer sensors from the eye tracker. We also use the gathered eye tracking metrics to construct a model to classify subjects into expert/novice. We find statistically significant relations between head movement as well as pupil diameter and periods of confusion. We identify a subset of the metrics that we use to build a simple classifier that is able to distinguish between novices and experts with accuracy of 84%.

A.3.1. Introduction

Arthroscopic shoulder surgery is an advanced orthopedic surgical procedure, which is particularly challenging due to the complex anatomy of the shoulder, and tight spaces for navigation, which also limits the view from the arthroscope. It is used to treat a number of disorders such as repair of torn tendons and rectifying chronic dislocation, as well as for diagnosis. In all of these procedures, the ability to quickly and effectively navigate through the joint to reach the desired location is essential. An important aspect of navigation is the ability to quickly recognize anatomical landmarks and to focus attention on the appropriate region of the arthroscope image. For assessment and training it is important to have an objective assessment of such perceptual and attentional aspects and to detect portions of the procedure where students may become confused. In this paper, we report on the results of the first cadaver-based study to analyze and compare expert and novice eye movement patterns in performing arthroscopic surgery. We study the diagnostic arthroscopic shoulder surgery task since it involves navigating to various parts of

the shoulder and inspecting them and thus allows us to focus purely on navigation skills. The existing studies on comparing eye movement patterns between experienced surgeons and novices have predominantly used VR training simulators [69], [150], [152], [155]–[157], still images of the surgery [99], [158] or physical box trainers [112]. We use so-called soft cadavers, which are specially prepared so as to retain the natural tissue properties. This means that our study is able to capture important aspects of the surgery such as tactile feedback and surgical setup not captured by simulations. Our work is also the first to study arthroscopic shoulder surgery. Previous eye-tracking studies of surgery have concentrated predominantly on laparoscopic surgery which usually involves anatomy of the abdomen. In contrast, the diagnosis of the shoulder requires the surgeon to navigate the arthroscope through bones and muscles inside the rounded shoulder joint. Experts can usually smoothly maneuver the arthroscope instruments with the automaticity developed through experience. In contrast, novices often experience confusion in trying to locate the anatomical landmarks from the magnified view of the operating site on the arthroscope output. Previous studies in the area of human-computer interfaces and intelligent tutoring have found pupil size and head movement to be associated with periods of confusion [159], [160]. We sought to determine whether these metrics can also be used to detect confusion during shoulder arthroscopy and found positive relationships between both and novice states of confusion. Ours is the first study to attempt to use objective metrics to detect confusion during surgery. An effective assessment instrument should be able to distinguish between performance of subjects with varying levels of experience and expertise. We thus analyze the differences in gaze metrics between experts and two groups of novices of varying experience. We identify a small subset of the metrics with good discriminatory power and use them to build a simple classifier that is able to distinguish between novices and experts with high accuracy. This leads us to conclude that there are significant differences in perceptual parameters between novices and experts in arthroscopic surgery that could be used for objective assessment as well as tutoring

A.3.2. Related Work

Arthroscopic skills are difficult to acquire because they require use of multiple tools, using both hands while viewing the surgical site on a two-dimensional display, with constant vigilance to the operating environment [161]. Arthroscopic surgery is taught as a core component in a majority of orthopedic residency programs. Cadavers are often the first choice of surgeons for practice because they provide a real anatomical experience [162]. Other methods that have been tested with varying success in orthopedic teaching include interactive computer simulation [163], physical simulation environments [164] and virtual reality simula-

tors [165], [166]. Approaches in assessing arthroscopic surgical skills include Global Rating Scales [167], motion analysis [168], virtual reality simulators [165], [166], and simple bench model arthroscopic simulators [169].

Eye tracking studies comparing experts and novices have been carried out in a number of surgical domains. Tien et al. [113] compared the gaze behaviors of experts and junior surgeons during key stages of a live open inguinal hernia repair. They found that experts have a higher fixation frequency and concluded that it could be due to lower mental demand resulting from automaticity developed through practice. Similar findings are reported by Erridge et al. [170] during live laparoscopic gastric bypass surgery. Novices were found to pay less attention to the operative site but more to the sterile field. A number of studies of eye movement patterns of experts and novices [69], [148], [149] found that experts tend to fixate on the target more often than the instruments. Meanwhile, Law et al. [149] reported that novices either alternate their gaze between the target and instruments, focus on objects in between the target and the instruments, or follow the instrument on its way to the target. A study by Hermens et al. [148] also found differences in eye movement statistics between experts and novices. The experts in their study reportedly had lower saccadic rates and higher peak velocity, independent of where these eye movements were aimed. Similarly, in a study of global eye movement parameters of expert and non-expert participants, Kocak et al. [112] found that experts had significantly lower saccade rates and higher peak velocity than non-experts. Beyond analysis of eye movement metrics, a number of studies have used the metrics to build models to classify subjects into expert and novice. Eye metrics and tool motion data have been considered as features in assessing the skill of a surgeon while performing functional endoscopic sinus surgery [153]. Hidden Markov models were built for seven different surgeries in two levels of expertise using the eye-gaze locations and the surgical tools motions. The findings revealed that eye-gaze data contains the skill-related structures, and combining it with the surgical tool motion data improves the classifier performance. Richstone et al. [114] used eye movement metrics to develop models to classify surgeons into experts and non-experts. In a simulated surgery they achieved 91.9% and 92.9% accuracy with the linear discriminant analysis and neural network analysis, respectively and 81.0% and 90.7% accuracy in a live operating room setting. Eivazi et al. [22] used a random forest classifier to classify micro-surgeons in the cutting and suturing tasks and achieved a 70% recognition rate for the detection of expert and novice groups. Rose and Pedowitz [171] investigate the assessment of basic arthroscopy skills using virtual reality modules developed through task deconstruction. Participants with the most arthroscopic experience performed better and were more consistent than novices on all 3 virtual reality modules. Greater arthroscopic experience correlates with more symmetry of ambidextrous performance. While no work has investigated detection of confusion during surgery, detection of cognitive affective states such as confusion and boredom has been studied in the field

of Intelligent Tutoring Systems. Pachman and colleagues [92] used eye tracking for early detection of confusion in a digital learning environment. In their study, the participants were asked to solve problems while their eye trajectories were recorded and this data was triangulated with self-ratings of confusion and cued retrospective verbal reports. Delucia and colleagues [172] sought to determine whether eye movements reflect confusion while users completed tasks with two simulated devices. They measured confusion using a subjective Likert measure in which subjects were asked to rate their agreement with the statement "I was confused" and were not able to find consistent common correlation patterns between the variables for both devices, but they found that higher confusion ratings were positively correlated with the total fixation time on the whole screen, mean fixation duration and task completion time. Lallé and colleagues [173] included pupil diameter and head distance to the target as the predictors of the user's confusion. They studied various combinations of gaze, pupil diameter, head distance and mouse events as predictors. The authors concluded that features of pupil size are strong predictors of confusion, which is consistent with the fact that pupil size is correlated with cognitive load, which plausibly correlates with confusion.

A.3.3. Participants, Materials and Methods

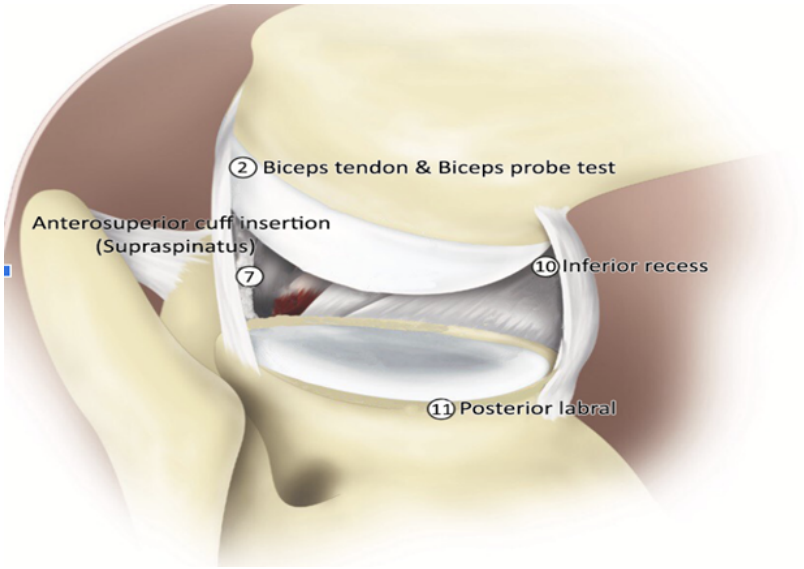

Figure A.12.: Portion of the shoulder anatomy with Landmarks 2, 7, 10, and 11

After obtaining approval from the Mahidol University Institutional Review Board, a total of thirteen participants (4 Females) were recruited. They consisted of four fellows (two to ten years of experience) from the Department of Orthopaedics,

Faculty of Medicine Ramathibodi Hospital, Mahidol University, and nine residents from the Orthopaedic Surgery Residency Program there. Five of the residents were in the third year and four in the fourth year. The residents were at an early stage of orthopedic training and were without prior arthroscopy experience. All the participants had normal or corrected-to-normal vision. Eye gaze data was recorded using the Tobii Pro eye tracker (Tobii Glasses 2.0, Tobii Sweden), which was calibrated by looking at a marker placed near the arthroscopic output screen. The cadaver (Male, 52 years old) was set up in the beach-chair position. An expert surgeon prepared the arthroscope setup (ConMed Linvatec) and inserted the primary portals into the shoulder prior to the procedure. The arthroscope camera output was displayed on a 52-inch screen which was placed four feet away from the participant. ARUCO markers were also placed around the screen in order to identify the screen in a later stage. Each participant was first acquainted with the cadaver setup, the diagnostic shoulder arthroscopy steps, and the evaluation study protocol. Each participant was asked to navigate and diagnose twelve anatomical landmarks within the shoulder in sequence (Table A.9). The portion of the shoulder anatomy from viewing with the scope in the posterior portal and four visible landmarks 2, 7, 10 and 11 are shown in Figure A.12. Among them, some are easy to navigate to and diagnose while some are more difficult. The landmarks which are categorized by the expert as hard to diagnose are highlighted and explanations are provided in Table A.9. For each landmark, the expert provided explicit verbal instructions with the name of the landmark (e.g. “Start Biceps tendon”) to navigate to and upon arrival at the landmark, the expert called out its name (e.g. “reached Biceps tendon”). The start and end times for each landmark navigation task were recorded as part of the data stream. Throughout the procedure, a think-aloud protocol was used and the participants were asked to describe their immediate objective, actions and any points at which they became confused (when they could not find the landmark or they did not recognize the part of the anatomy they were in). In addition to the self-reported confusion, a member of the investigation team also monitored the participants and recorded portions of the performance as confusion in situations when a participant paused or made non-goal directed movements for a period of time which was followed by the attending surgeon’s assisting intervention. The study spanned two days, with the left shoulder of the cadaver used on the first day for six participants, and the right shoulder used on the second day for eight participants.

Data Preparation

From a preliminary study, we found that while surgeons diagnose a landmark, they tend to look at the center of the scope image and tend to look at the area near the circumference of the scope image in the direction of the next landmark to visit

Anatomical Landmarks	
1.	Rotator interval
2.	Biceps tendon & Biceps probe test: easy to find long head biceps (LHB) but difficult for use probe to handle LHB (need another hand to control the probe)
3.	Biceps anchor
4.	Labral superior to anterior
5.	IGHL
6.	Subscapularis tendon and insertion
7.	Anterosuperior cuff insertion (Supraspinatus)
8.	Posterosuperior cuff insertion (Infraspinatus): difficult to move from supraspinatus to infraspinatus (need to control the camera backward along the tendon).
9.	Bare area
10.	Inferior recess: difficult move from the posterior chamber downward direction to the inferior chamber
11.	Posterior labral: difficult to slide the camera from inferior chamber to posterior than to superior chamber (the camera could easily back out from the trocar due to the limited space)

Table A.9.: Twelve anatomical landmarks to diagnose (The landmarks which are categorized by the expert as hard to diagnose are highlighted.)

before moving the scope. We, therefore, define four areas of interest (AOIs): the center area of the scope image (the inner circle) (Figure A.13), the outer area of the scope image (outer circle) (Figure A.13), the arthroscope output screen (outside of the scope image), and the shoulder area on the cadaver (Figure A.14). Eye-tracking metrics considered in this study are the rate and duration of fixations/saccades, the time to first fixation and the duration of the first fixation were calculated with the Tobii-I-VT Attention Filter using default parameters. Fixation is the visual gaze on a single location and saccades are the rapid movements of the eyes that abruptly change the point of fixation. The field view videos of the eye tracker were processed to demark the AOI's. The arthroscope output screen was detected using ARUCO markers and the scope view on the screen was detected

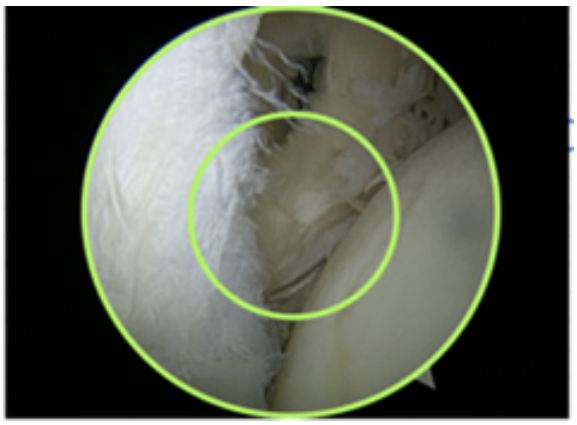

Figure A.13.: Inner and outer circles on the scope output.

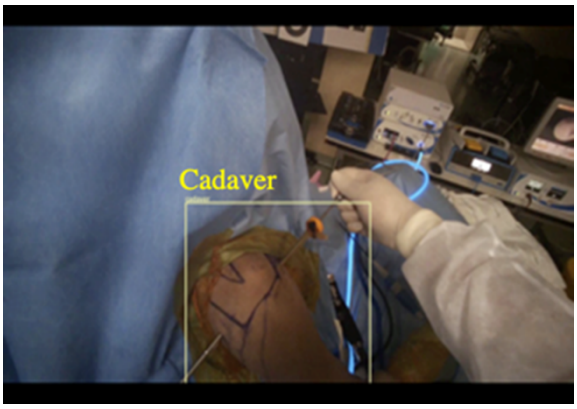

Figure A.14.: Detected cadaver shoulder on the video.

using a simple circle detection method (cv2.circle()). The cadaver area in the video frames was detected by using the YOLOV3 CNN object detection model [174] trained using transfer learning. The cadaver shoulder in the video frames was labeled using the video labeler app from Matlab (R2019b). We used the video frames from three participants for the left shoulder and from two participants for the right shoulder area.

A.3.4. Analysis and Discussion

The two most commonly studied features of eye movement are fixations and saccades. Fixations are visual gazes on a single location whereas saccades are rapid eye movements between fixations. Among the large number of possible eye tracking metrics, those commonly used in medical studies are fixation rate (number of

fixations per second), saccade rate (number of saccades per second), fixation duration (length of each fixation), saccade duration, average time to first fixation, and duration of first fixation [112]–[116]. We thus chose these metrics for the current study. Along with the eye metrics, we used the completion time as an objective measure of skill. We categorized participants into three groups: four experts as E, four third-year residents RY3, and five fourth-year residents as RY4.

Gaze Data Analysis

As shown in Table A.10, the average fixation rate of experts is higher than novices, but the expert's average fixation duration is the lowest among all the groups. The average saccade rate and duration (ms) of experts is higher than the RY4 group. The expert's average time to the first fixation is the lowest among the three groups, the average fixation duration is less than that of RY4.

Eye Gaze Metrics

	Expert	**RY3**	**RY4**
Avg. fixation rate	3.01	1.62	1.93
Avg. saccade rate	0.71	0.39	0.82
Avg. fixation duration (ms)	411.24	490.37	466.33
Avg. saccade duration (ms)	29.90	35.77	28.24
Avg. time to first fixation (ms)	50.00	155.00	450.00
Avg first fixation duration (ms)	1,039.50	499.80	1,269.25

Table A.10.: Eye gaze metrics.

Overall, experts have higher fixation rates compared to the novices and the majority of their fixations fell on the scope image. To investigate the fixation patterns of the expert and novice in the inner and outer circles AOIs of the scope, we considered 80% of the process of navigating from one landmark to another into finding the general area of the landmark and another 20% as zeroing in on the landmark. We found that during the 80% portion experts and novices both tended to fixate more on the outer circle in a ratio of roughly 2:1. During the 20% portion the experts fixated on the inner circle with a ratio of 2:1 while the novices continued

to fixate on the outer circle with roughly the same ratio as before. This shows that the experts adjust their focus of attention to suit the portion of the navigation task, while the novices keep their focus primarily in only one area. This could be explained by the fact that an expert would be expected to know that they are getting close to a landmark whereas a novice might not.

Confusion

With a handful of reference anatomical regions within the joint, novices often miss the target landmark to diagnose during the procedure. Failure to recognize landmarks may result in disorientation and confusion as a student seeks to navigate through the shoulder joint. Since previous studies in user interfaces and intelligent tutoring had identified significant relationships between user confusion and metrics of pupil diameter and head movement, we sought to determine whether such relationships exist in this surgical domain as well. As head movement metrics, we used the gyroscope and accelerometer data available from the Tobii eye tracker. Six novice participants (3 RY3, 3 RY4) reported a total of 14 confusion points while navigating and diagnosing at landmarks 1, 3, 6, 7, 8, and 12. The number of confusion points per landmark ranged from one to five with the highest frequency of three times reported at the landmarks 1, 6 and 8.

The follow-up interviews with the experts revealed that novices might get confused in landmark 1 due to a lack of recall of the background knowledge. At landmark 1, instead of looking for the void triangular space of the rotator interval between the subscapularis and glenoid and supraspinatus, the novices tended to look at the nearby structure. While in landmark 6, the novices need to locate the insertion of supraspinatus on the humerus. In the experts' opinion, the novices mostly focus on the tendon part, while all experts specifically focus on the tendon insertion point. This may be related to the level of knowledge of the pathological area on this tendon. The infraspinatus at landmark 8 is a tendon posterior to supraspinatus tendon. These tendons are blended together and have the same texture. Therefore, the location of infraspinatus can be identified only by understanding the exact location of infraspinatus (posterior half of these blended tendons).

In terms of the time taken to complete the task, the experts completed the task with the least amount of time to diagnose at each landmark and had the least variation in task times. We observe that some landmarks require more time to navigate to and diagnose, particularly landmark 2 and 6 which are categorized as hard to diagnose. On average, the six novices who became confused took 1.5 times and 2 times longer than other novices in hard and easy landmarks, respectively.

The Percentage Change in Pupil Diameter (PCPD) is an objective measure of cognitive skills. Kruger et al. [111] studied PCPD as a measure of cognitive load and compared it with different cognitive load metrics including EEG, heart rate and

blink rate when students were watching a recorded academic lecture, with and without subtitles. They found that higher cognitive loads were associated with higher PCPD values. We expect that the subject's cognitive load will increase while navigating the arthroscope in the landmarks where confusion was recorded. To determine that, we need a period of low cognitive load as a baseline. We used the period from the end of the previous landmark until the beginning of the current (confused) landmark as the baseline period since during that period the subject just is not actively navigating through the joint. The PCPD value was computed by subtracting the average diameter from the (confusion) landmark from the baseline diameter and divided it by the baseline diameter. From the six participants who became confused, the PCPD ranged from a minimum of 0.91% (left eye) and 0.97% (right eye) to a maximum of 1.22% (left eye) and 1.12% (right eye). On average, during the periods of confusion the pupil diameter changed by 1.02% in the left eye and 1.03% in the right eye relative to the baseline. The minimum values came from two novices at five different landmarks; all others had positive change in PCPD.

We investigated the head movement of the novice participants during the landmarks with confusion using the information from the gyroscope and accelerometer sensors of the eye tracker. Confusion was not reported in landmark 2 (L2: Biceps tendon & Biceps probe test) for any of the novices and hence it was considered as the baseline. We compared the head rotation and acceleration information between novices with and without reported confusion by computing the differences between the minimum and maximum values in x-, y- and z-axes. T he differences are compared with the baselines using a paired t-test for each participant with confusion reported. The differences are significant in all three axes for head movements from the accelerometer as well as in y- and z-axes from the gyroscope sensors (p-value = 0.05). As shown in Table A.11, the average differences between the two groups are substantial in the x-axis for head rotations and the z-axis for acceleration.

Figure A.15 the rotation from the gyroscope sensor and A.16 shows the acceleration from the accelerometer along the x, y, z axes of a novice participant (RY4). As shown in the figures, this particular novice rotates the head along the x-axis and moves along the z-axis while navigating the arthroscope to the landmark 6 and performing the insertion (Subscapularis tendon and insertion).

Classification

In order to evaluate whether the eye-tracking metrics can be used to assess level of expertise in arthroscopic shoulder surgery, we sought to build models to classify participants as novice or expert. Due to the small size of the data set, we used leave-one (participant)-out to validate the classifiers. We applied Synthetic

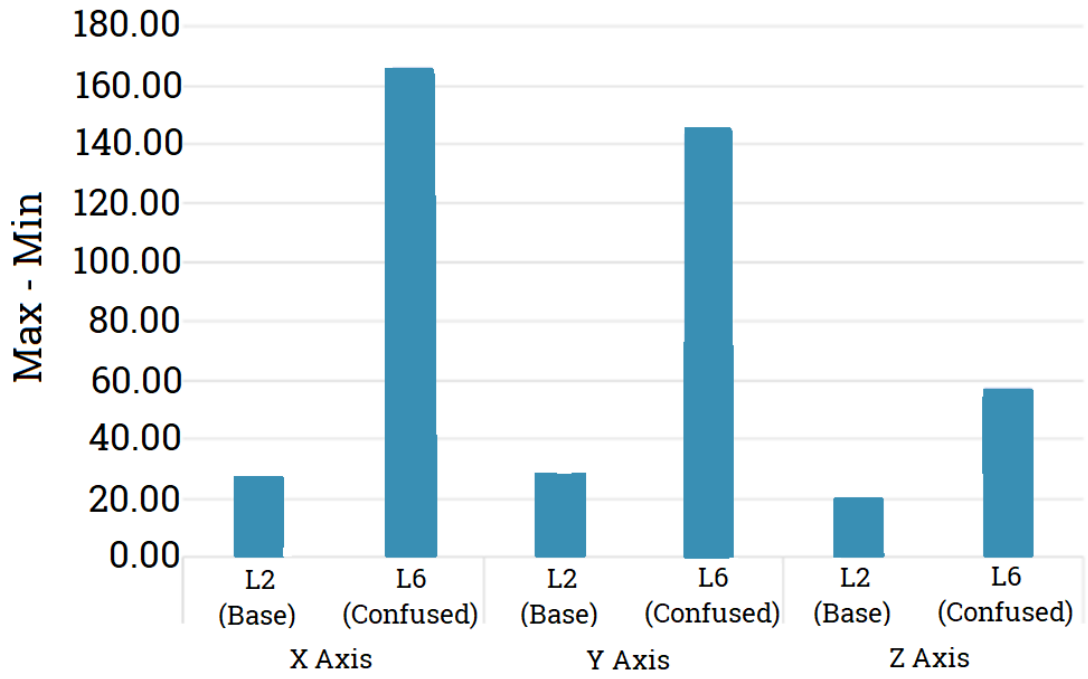

Figure A.15.: (a) The differences between minimum and maximum in three axes at the baseline landmark (L2) and the landmark with confusion (L6) from the gyroscope sensor.

Minority Over-sampling Technique (SMOTE) repeatedly to the remaining twelve participants' gaze features. In each iteration, we randomly selected three novices and four experts, and generated one instance of novice with SMOTE and added it back to the novice data pool. The process was repeated until we reached a total of 100 novices. In the same manner, we generated expert data instances until we achieved 100, resulting in a balanced data set with 200 instances. Features considered for the classification model included twelve gaze metrics extracted from the eye data including fixation and saccade rates for the whole procedure and three AOIs, average fixation and saccade rates, time to first fixation, and duration of first fixation. We selected the best five features using the information gain ratio (Table A.12). With the logistic regression model, we achieved a classification accuracy of 84%. The logistic regression model misclassified an expert and an R3 novice who have similar fixation rates (gaze points/sec) and an R3 novice with similar time to first fixation with an expert. The results show that in the domain of arthroscopic shoulder surgery, although the differences in eye-movement data are multidimensional, the two groups of participants can be classified with high accuracy by a simple model.

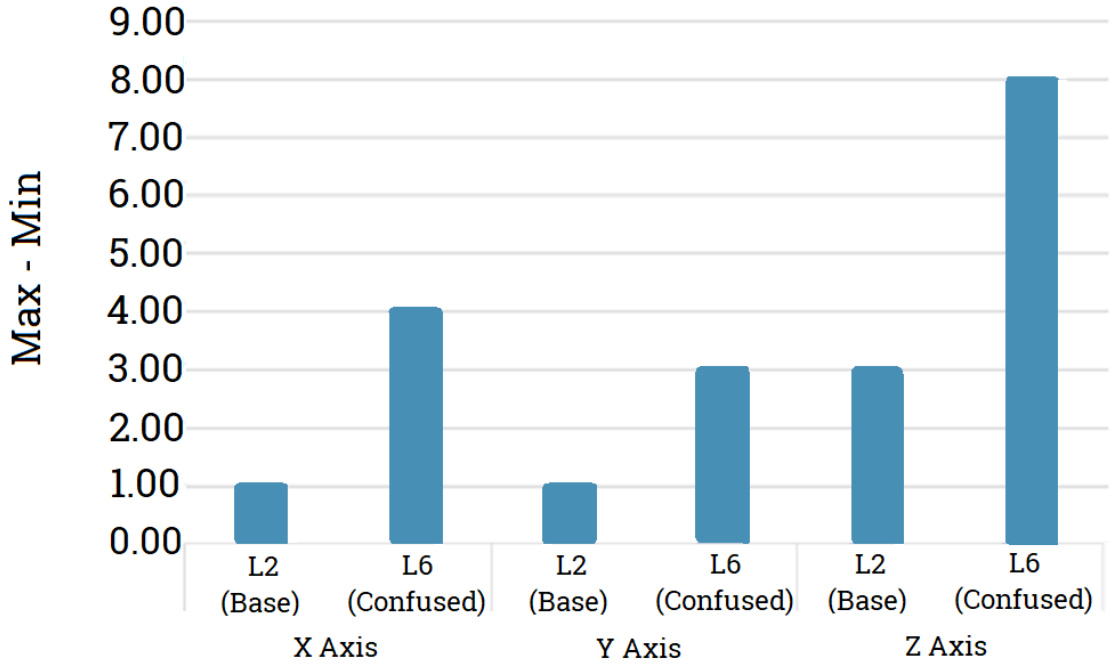

Figure A.16.: (b) The differences between minimum and maximum in three axes at the baseline landmark (L2) and the landmark with confusion (L6) from the accelerometer sensor

A.3.5. Conclusion

The required skill set for arthroscopy is complex, due to an indirect view of the surgical site through the arthroscope, limited tactile feedback, and complex hand-eye-coordination. The operative time, probe path length, and number of movements are commonly utilized as surrogate markers for assessing skills. While previous studies have centered around the dexterous aspects of motor skills, we investigate cognitive aspects by studying the differences in perception between participants of differing experience [165]. During the arthroscopic surgery, surgeons rely primarily on visual information. Perception and attention are two separate but related processes. Initially attention occurs, and perception follows. This study has shown that there are significant differences between expert and novice focus of attention during the arthroscopic navigation task both overall and during particular portions of navigation. We investigated a number of other questions such as the relationship between user confusion and metrics of pupil diameter and head movement, as well as whether the eye-tracking metrics can be used to classify the experts and novices. In contrast to the existing studies, the gaze measures in our study are

Gyroscope and Accelerometer

Sensor		x – axis	y– axis	z – axis
Gyroscope	Novices with confusion	106.67	52.50	28.17
	Novices without confusion	22.93	30.12	13.08
Accelerometer	Novices with confusion	2.00	2.50	4.67
	Novices without confusion	1.37	0.99	1.83

Table A.11.: Comparison in average differences in minimum and maximum values in three axes between novices with confusion reported, novices without confusion

Selected Features

Feature	Gain Ratio	Min, Max, Mean
Time 1st fixation (ms)	0.482	Expert: 25.0,75.0, 69.2 Novice: 75.0, 1075.0, 181.2
Fixation Rate	0.418	Expert: 2.0, 3.6, 3.3 Novice: 1.2, 2.6, 1.6
Fixation Rate AOI In	0.381	Expert: 9.6, 15.0, 12.0 Novice: 5.9, 13.3, 10.0
Avg Fixation Dur. (ms)	0.358	Expert: 221.7, 705.9, 302.4 Novice: 354.2, 640.8, 489.4
Avg Saccade Dur. (ms)	0.306	Expert: 26.0, 35.0, 28.4 Novice: 25.4, 42.9, 34.2

Table A.12.: Selected features with the information gain ratio.

collected with the cadaver specimens which provide the most realistic experience. We have demonstrated the potential of eye-tracking to provide reliable tools for automatic performance assessment in arthroscopic shoulder surgery. This leads us to the conclusion that gaze data carries important information about the skills of arthroscopic surgeons which could contribute to automated objective assessment. The future steps of this research include the development of an intelligent training system in the virtual reality environment that dynamically detects novice confusion and classifies surgeon's performance based on eye-movement data

Acknowledgement

This work was partially supported through a fellowship from the Hanse-Wissenschaftskolleg Institute for Advanced Study (HWK), Delmenhorst, Germany to Su Yin for collaborative work with the University of Bremen, and through a study group grant from HWK to Haddawy. It was also partially supported through a grant from the Mahidol University Office of International Relations to the MIRU joint unit.

B. Cross-Domain Generalization

This chapter is based on the following publications:

- **B.W. Hosp**, F. Schultz, O. Höner, and E. Kasneci. "In Search of A Superior Gaze Behavior: Cross-Domain Shared Expertise-Related Gaze Features."

- **B. W. Hosp**, F. Schultz, E. Kasneci, and O. Höner. "Expertise classification of soccer goalkeepers in highly dynamic decision tasks: A deep learning approach for temporal and spatial feature recognition of fixation image patch sequences," Frontiers in Sports and Active Living, vol. 3, p. 183, 2021.

B.1. In Search of A Superior Gaze Behavior:Cross-Domain Shared Expertise-Related Gaze Features.

Abstract

When we talk about perceptional expertise, we usually talk about it in certain limits like a domain or a task. However, so far, there has been no proof found that states expertise is restricted to such limits. Perceptional expertise might also have some kind of domain- or task-independent source, which is shared by experts from different domains. Such perceptional expertise is considered to prove that it is possible to generalize gaze behavior and describe it as a domain-independent skill. Seeing generalized, cross-domain perceptual expertise definition as a far-reaching aim, a first step is to find commonalities and differences between experts from different fields. Therefore, we are investigating a minimal set of features from one domain to build a machine learning model and predict the expertise of samples from two other domains. The diversity of the performance values might indicate that not domain but task similarity or other boundary conditions are more important for generalization.

B.1.1. Introduction

On the one hand, it is assumed that experts develop their optimal methods of perception by solving highly similar tasks for many years and optimizing their perception in the process. Thus, expertise forms over years of experience and practice. On the other hand, however, it is assumed that there are certain commonalities in the gaze behavior of experts. In addition to these commonalities, the differences between levels of expertise are also of particular interest for research [43]–[49], [67]. This interest originates from the ability to derive findings of perception at different developmental stages, but also from the ability to develop the diagnostics as a foundation for possible support options, based on findings of perception research. Different expertise classes show different similarities in perception so that a beginner needs completely different assistance than an advanced user [41]. In recent years, perception of experts has been investigated in various fields and tasks [13]–[15], [41], [100]. In sports psychology, expertise has been linked to more efficient gaze behavior in decision-making tasks [44]–[49], [67]. However, aspects of perception, that allowed separation of expertise classes, were often found, but could not lead to consistent results. Thus, findings are often dependent on domain and task type. While a look at the current research situation shows a mass of expertise research studies, only little inter-domain or inter-task work is done, so most work is somehow limited to a task or domain. Gegenfurtner et al. [72] show

that it is possible to transfer expertise from familiar tasks to semi-familiar tasks, but not to unfamiliar tasks. Likewise, they took the same subjects for both tasks, which introduces a high risk of enabling recognition of subject-specific characteristics instead of expertise. Thus, while differences have often been found, only little is known about inter-task or at least inter-domain expertise that is transferable or generalizable. The problem of a missing generalizable feature set that works for more than one task or domain, has yet not been confirmed. So far, no dedicated set of traits has been found that is better suited to recognize expertise than others. Therefore, previous study results could hardly or not at all be transferred to other studies and were always limited to one field, task or at least data set [42]. However, since it is expected that experts in the same task exhibit certain commonalities in gaze behavior, in a subsequent step, experts could also exhibit certain commonalities regardless of the task or even domain. To prove this hypothesis, studies are needed that evaluate the gaze behavior with the same methods, on different tasks, or in different domains. The overall question is whether expertise-related features derived from visual behavior are consistent across domains and whether experts from different domains share (at least some) visual strategy features. A superior set of perceptual properties would lead to a complete overturning of our understanding of expertise. Such perceptional expertise is considered to prove that it is possible to generalize gaze behavior and describe it as a domain-independent skill. Seeing generalized, cross-domain perceptual expertise as a far-reaching aim, a first step is to find commonalities and differences between experts from different fields. Therefore, we are investigating a set of features, shared by three different domains. We use a minimal feature set to infer expertise classes by training a machine learning model with data of one domain and testing the model with unknown data from the two other domains, by predicting the expertise classes of the new data.

Especially, when looking at highly dynamic tasks from a more generous perspective, one can see that it typically consists of fast movements. Decisions need to be made in little time and have usually a high impact on the continuation of a task. To capture underlying cognitive processes with eye tracking, we use features from even volatile and fast movements recordings of the eyes.

B.1.2. Methods

In the first step, we collected all the gaze data from three distinct studies that we conducted. Each of the data sets contains samples of subjects that were previously assigned (based on their skill) to one of the following classes: expert, intermediate, or novice. As we use supervised learning algorithms, we need an external classification to label some of the samples we collected as belonging to the correct class. We do this for a certain amount of samples but equally distributed on each

of the present classes. This bunch of data is called the training data set. With such data, we train our algorithm to recognize the connection between a training sample (defined representation of the gaze behavior in form of samples for each video, operation, image, etc.) and its correct class (novice, intermediate, and expert). By training the algorithm, we want to use a representation of the gaze data that optimally describes the gaze behavior of a sample. The better the representation can describe the commonalities of samples from the same class and differences of the samples from different classes, the better our algorithm can be taught how to predict the class of new, unknown samples. In the next step, some samples, that have not been labeled yet are fed to the model. The model has no idea about the class membership of this bunch of samples, which is called the testing data set. By feeding the model this unknown data, we can estimate how well the model behaves when we collect more data and predict their class membership. It is therefore a quality measure. We applied this method to all three of the following studies. Thus, we build a model, that 1) can recognize different skill levels of subjects based on their gaze behavior, 2) prove how well it behaves on unknown data, 3) allows the whole process to be reproducible and objective and 4) if possible, provides insights about the perceptional development between the expertise classes in a cross-domain manner.

Study A

Data set A contains the samples of 33 soccer goalkeepers and 28 soccer field players from two studies in virtual reality on decision-making. For these, we took an HTC Vive with an integrated SMI eye tracker, which is capable of recording the eyes with 250 Hz. We defined typical in-game scenarios of soccer. Resulting in unique videos, in which youth players of the VfB Stuttgart played the defined scenes on the training space of the youth performance center of the VfB Stuttgart. While they were acting the scene, a 360° camera was placed at the position of the subject to capture the realistic field of view. The task of the subjects was to decide how to continue the game after the last return pass to the position of the subject on the field, as the screen went black after the last pass. We collected data of n=12 experts (expert youth soccer goalkeepers from U-15 to U-21) during two youth elite goalkeeper camps of the DFB. Data of the n=10 intermediate players were recorded in our lab. The intermediates are goalkeepers from the regional league in Germany (semi-professional). The novices (n=13) had no experience in competitions and no training on a weekly basis, but up to 2 years of experience. The study was confirmed by the Ethics Committee of the Faculty of Economics and Social Sciences of the University of Tübingen.

The subjects of the field player study (n=14) are all from the VfB Stuttgart youth elite program. Therefore, they are all considered to be experts. They all play higher than the regional league and have a lot of experience in competitions.

Study B

The second study was made in arthroscopic surgery, where, usually everything takes place in front of the surgeon. As such, a field of view camera provides a much better resolution of a smaller area, which is technically better suited. Especially in arthroscopic surgery, the main focus of the surgeon lays on the patient and the scope output which is usually a big screen in front of the surgeon where the arthroscopic camera sends its video feed. With such a camera, more details can be captured. We asked surgeons to navigate an arthroscope through a portal on a soft cadavers' shoulder to the operating site where the tendon of the shoulder needs to be repaired. The surgeons were standing in front of the soft-cadaver and 4 feet further away we placed a 4k, 52-inch screen which showed the output of the arthroscope. We gave surgeons a head-mounted eye tracker (Tobii Glasses 2) during that arthroscopic surgery, namely a shoulder tendon repair operation. The study was approved by the Mahidol University Institutional Review Board. We captured the data of n=15 subjects in an operating room of the Ramathibodi Hospital of the Mahidol University in Bangkok, Thailand. The expert group (n=5) are fellow surgeons from the Orthopaedics Faculty of Medicine of the Mahidol University, who have 4-10 years of experience in arthroscopic surgery. A second group, we now and later call intermediates, consists of n=5 surgeons being in their fourth year of the Orthopaedic Residency program. In the last group, we call novices, we collected data of n=5 surgeons being in their third year of the Orthopaedic Residency program. The intermediates as well as the novices had no experience in arthroscopic surgery before. The difference between intermediates and novices is mainly based on the one year of medical education between them.

Study C

The data of study C is coming from a more static task. In study C we collected data of 58 dentists during OPT analysis. N=17 subjects are novices. On recording day, they have been in their 6th semester of dental studies before their first course in OPT reading. The intermediates (n= 14) have been in their 10th semester. Thus, they have more experience than the novices and already visited two courses in OPT reading. The dental experts are dental physicians who have already practiced for several years in their field. All subjects from this study are students from the University of Tübingen and/or are working at the University Hospital Tübingen. The task for the dentists was to mark anomalies in multiple radiographs. Thus, the task was quite static. Therefore, we limited the time for each radiography to be marked by a dentist. This leads to a more dynamic gaze behavior. The data of the dentists were captured with an SMI RED 250 remote eye-tracker which was attached to a common laptop. The whole procedure of marking anomalies has been performed on such laptops.

Procedure

The experts in all data sets are classified based on either years of experience in the task or being picked by talent scouts. The novices are defined as beginners of the field or having no experience in the task. The intermediates are loosely defined as in between, with more experience than novices but way less than experts. As all data sets were captured with a different eye-tracking device, we first looked at all the features from all data sets and defined a subset of features that are shared by all of the data sets. In the next step, we split the data of experts, intermediates, and novices in each data set. We defined a balanced training set of a randomly picked data set and trained a bagged tree model. With this first model, we used an MRMR technique for feature selection. Subsequently, we ranked the features by their importance for the model during cross-validation. With the new subset of features, we now used data from the two remaining data sets to test the model on other domains. In the following, we will talk about our observations.

B.1.3. Results

After the first feature ranking, we end up with a sub-set of features that has the highest impact on accuracy. We, therefore, pick them as candidates for a subset of features that are shared by all data sets. The most promising subset of features was the following:

- maximum saccade peak velocity
- maximum fixation dispersion
- standard deviation of saccade peak velocity
- maximum saccade amplitude
- minimum smooth pursuit dispersion

With these features, we were able to achieve an accuracy performance of 58%. This sounds quite low, but we need to remember that we are looking at a three-class problem. Thus, the chance level of picking the right class is 33.33%. With 58% we are slightly worse than the doubled chance level. An accuracy of over 66% would lead to the fact, that single samples might be classified incorrectly, but the majority is classified correctly. Therefore, also the majority of a subject's samples are classified correctly and subsequently the subject in total, too. Looking at the two data sets that were classified, the dentist data set (study C), had a total classification accuracy of only 29%. The intermediates were classified with 7.7%, the experts with 35%, and the novices with 45 %. Therefore, the dentist

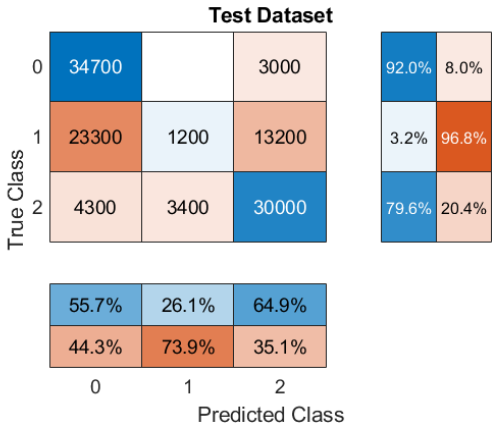

Figure B.1.: Confusion matrix showing predictions with data of new data set. Class 0 = novices, class 1 = intermediates, and class 2 = experts.

data set is slightly worse than the chance level and thus, we might have not the most optimal features for that data set. Another reason for this classification might also be the different task of static diagnostics. On the soccer data set (study A), which task was much more similar to the surgeons, we reached an accuracy for the novices of 83.4%, 0.8% for the intermediates, and 98.4% for the experts. Again, because of the miss classifications of the intermediates, the average accuracy is at 60%. With the mentioned features we were able to train a model with one data set and classify the two other data sets with an accuracy of 58%. On a deeper look at the single classes, we can see that the novices (92.0%, see Figure B.1) were nearly optimally detected, the intermediates with 3.2 % not at all, and the experts still with an accuracy of over 79.6%. From 100 runs 34,700 samples were correctly classified as novice and 3,000 falsely as an expert. This is not a problem, as we know in expertise research there are subjects acting better than their initial classification. More problematic is the amount of samples that belong to the expert class but is classified as novice or intermediate. In 100 runs 30,000 samples were classified correctly as expert samples. 4,300 samples incorrectly as a novice, and 3,400 samples as intermediate. At first, these results look complex to understand, but a closer look at how the intermediates are defined reveals the ambivalence of these results and a weak point in the classification. This will be discussed in the discussion part of this paper.

Shared, latent expertise features

Having a deeper look at the features and their characteristics, we can see three important correlations. As we normalized the data based on their data set, the values can be positive as well as negative. For comparison, this is important, as

the correlations are only visible there. The surgeons' experts e.g. have a maximum saccade peak velocity of -221.560 $^\circ/s$, followed by the intermediates with -0.7267 $^\circ/s$ and the novices with 222.287$^\circ/s$. Comparing the values with those of the soccer players, we see that the experts also have a highly negative value of -593.31 $^\circ/s$ followed by a high value of 234.56 $^\circ/s$ and an even higher value of 1211.377 $^\circ/s$ for the novices. Soccer players have more or less the same trend between the expertise classes. In the data set of the dentists, we cannot see this trend. Only experts and novices show similar values, thus, intermediates will be miss classified as novices (their values correspond much closer to the novices). For the dentists, we found a correlation between the trends of the standard deviation of the peak velocity. The dentists, as well as the surgeons, follow the same trend (experts: ca. -15$^\circ/s$, intermediates: ca. 6.5 $^\circ/s$, and novices: ca. 10$^\circ/s$). Here the data of the soccer players do not fit at all. A feature whose values correlate with both other data sets' experts and novices, is the minimum smooth pursuit dispersion. The values for the expert groups are slightly positive (0.019 to 2.25 pixels), while the values of the novices are slightly negative (-4.8 to -0.15). Only, again, the soccer players' intermediates correlate with the surgeons by being negatively close to zero.

B.1.4. Discussion

In this paper, we defined a feature set that is thought to explain expertise in multiple data sets from different domains. We trained a machine learning model with one data set and predicted the classes of two other data sets. With an average accuracy of 58%, the total performance is quite sobering, but on a more detailed look, the performance value can be understood quite easily. The differences in the domains seem to be not important, as it is possible to detect novices in dentistry with 45% accuracy. Thus, there might be some kind of general gaze behavior explainable throughout the two domains. The detection of the expert dentists is 34% much lower and close to chance-level. Thus, there are differences between the two data sets that hinder a generalization. As the data set of soccer players show high-performance values, we cannot say that different domains need different gaze behaviors, as our model is trained with surgeons' gaze behavior and could predict novices and experts of soccer pretty well. A much more important difference than domain, are the boundary conditions that we need to take into account. A possibly important difference between the dentist and the other two data sets was their low dynamics. The dentists were observing images on a laptop screen while the surgeons as well as the soccer players were allowed to move their heads completely free. Likewise, surgeons and soccer players needed to gain an overview, navigate, and decide how to continue, while dentists only marked anomalies on a fixed image. Thus, we support the findings of Gegenfurtner et al. [72], that it is

possible to transfer knowledge about gaze behavior of one task to gaze behavior of a familiar-tasks, but not unfamiliar-task. To allow a statement in the direction of cross-domain expertise-related gaze features, there need to be more investigations with data sets from different domains and/or different tasks, but similar hardware setups (same eye tracker, same speed, etc), but with this current work we can state that to generalize expertise-related gaze behavior between different domains, the task seems to be much more important than the domain itself.

B.2. Expertise Classification of Soccer Goalkeepers in Highly-Dynamic Decision-Tasks: A Deep-Learning Approach for Temporal and Spatial Feature Recognition of Fixation Image Patch Sequences

Abstract

The focus of expertise research moves constantly forward and includes cognitive factors like visual information perception and processing. In highly dynamic tasks, such as decision-making in sports, these factors become more important in order to build a foundation for diagnostic systems and adaptive learning environments. Although most recent research focuses on behavioral features, the underlying cognitive mechanisms have been poorly understood, mainly due to a lack of adequate methods for the analysis of complex eye-tracking data that goes beyond aggregated fixations and saccades. There are no generally applicable statements about specific perceptual features that explain expertise. However, these mechanisms are an important part of expertise, especially in decision-making in sports games as highly trained perceptual-cognitive abilities can provide athletes with some advantage. We developed a deep learning approach that independently finds latent perceptual features in fixation image patches. It then derives expertise based solely on these fixation patches which encompass the gaze behavior of athletes in an elaborately implemented virtual reality setup. We present a CNN-BiLSTM-based model for expertise assessment in goalkeeper-specific decision tasks on initiating passes in build-up situations. The empirical validation demonstrated that our model has the ability to find valuable latent features that detect the expertise level of 33 athletes (novice, advanced, expert) with 73.11% accuracy. Our model is a first step in the direction of generalizable expertise recognition based on eye movements.

B.2.1. Introduction

In general, expertise research spans many different areas. Expertise research based on behavioral data has found its way especially into several fields, i.e. dentistry [30], surgery [20], [98], [175], and sports [63], [176]–[180]. In all of these areas, the assessment of user expertise is a fundamental task. By estimating the expertise of a user as accurately as possible, adaptive systems can be built to model different, distinct expertise classes and potentially create tasks specifically adapted to the expertise class. For diagnostics within the framework of sports science expertise research, groups of different performance levels are examined using the 'expert-novice paradigm' [181]. According to Tenenbaum et al. [182], this is the

most efficient way to study the development of cognitive and motor skills. Based on this paradigm, Ericsson et al. [43] developed the frequently used framework of the 'Expert Performance Approach'. This approach assumes that a subjects' behavior in a laboratory task is closest to their behavior on the pitch if the laboratory setting is as realistic as possible. It is therefore required to establish the highest possible ecological validity of laboratory tests, taking into account the internal validity [63]. According to this assumption, within the Expert Performance Approach, sports-specific scenes are often selected as stimuli for diagnostic [183]. However, in previous studies, the video stimuli were mostly presented on large screens or PC monitors and often from a third-person perspective (for review, see e.g. [67], [184]). This classical laboratory setting results in a low external validity [185] (for an overview see [186]). The trade-off of these validities plays an important role. Especially in highly dynamic environments, it is difficult to obtain robust and natural data. Robust data is obtained in highly controlled environments while natural data is obtained in natural environments. Therefore, these two aspects are opposites and relative to discussions about the tension between the internal and ecological validity of scientific studies. This is especially true in fields such as sports where besides, tactical and physical components, highly refined perceptual-cognitive abilities are key to success [61], [117]–[119]. Due to the fact that in high-level sports the physical strain of the athletes is significant due to intensive training schedules, enhancing cognitive factors like decision-making without additional physical training is gaining in importance [187]. For this reason, research efforts to identify the major cognitive factors leading to differences in performance, especially in regard to decision-making in the sports game, have increased in recent years. One aim of these efforts is the development of valid diagnostics that can, for example, identify the gaze behavior of experts engaged in successful decision-making. Accordingly, by teaching this gaze behavior it may be possible to design training programs that lead to improved decision making.

Due to ongoing technological development in the field of virtual reality (VR), it is now possible to present 360°stimuli from a first-person perspective in head-mounted displays (HMD). This increases the feeling of 'presence' for participants, which is defined as the psychological experience of 'being there' [188]. An increased feeling of presence should lead to more valid results as compared to presentations on screens [189], [190]. In addition to the valid stimulation and recording of behavior, an analysis of the underlying mechanisms of expertise is necessary to formulate explanatory approaches for identified performance differences. In recent years, cognitive processes (e.g., decision-making under pressure or anticipation of the continuation of a scene) in sports games have been studied. Thereby, new developments in image processing, measurement methods, machine learning, and eye tracking may be used to control the stimuli or utilized as non-invasive methods that do not influence the natural behavior of the athlete. The developments in eye tracking have shown that these methods of measurement hardly

disturb natural behavior, but, instead, become increasingly accurate and informative because cognitive processes like perception are very simple, non-invasive, and meaningful to track.

In sports science, the non-invasive method of eye tracking is considered a common and objective research method for the analysis of visual attention and the intake of visual information (for an overview see [191]). Here it is also assumed that the measurement of athlete gaze behavior in real sports situations generates the highest ecological validity. Mobile eye trackers have disadvantages (e.g. inaccurate measurements due to slippage, low frequencies), that can be circumvented by eye trackers integrated into the HMD. Due to the 360° videos that can be presented there, gaze behavior can be recorded at high frequency (up to 250 Hz) in ecologically valid environments with high experimental control.

The type of analysis also plays an important role because up until this point eye-tracking data has mainly been evaluated manually, visually or with statistical methods [108]. A newer and popular technique to classify expertise is to train a model by a brute-force approach of all possible features available from the data. Hosp et al. [176] use this technique to investigate the expertise of soccer goalkeepers by recording their gaze during the game build-up. In their approach to expertise recognition, they take all possible features provided by the eye-tracking vendor and add derived statistical features on top. They find a support vector model (SVM) with high accuracy. However, this feature crafting is highly time-consuming and does not necessarily provide the most suited features. There is no real evidence that certain features or feature combinations highlight expertise. Fixations, saccades, and their frequencies and lengths are often used, but can not lead to a full understanding of expertise as Klostermann and Moeinirad [42] revealed. They conclude that single features describing gaze behavior are only conditionally suitable to classify expertise differences or, at the very least, have yet to be found. Rather, expertise comes from the optimized perception of helpful gaze locations and the sequence of these locations also called scan path. To explore the gaze locations and their temporal succession, our approach is to let artificial intelligence (AI) describe the features around these gaze locations (albeit very abstract). In doing so, the AI itself decides which shapes, colors, corners, and edges in the fixation locations are considered important for distinguishing expertise. This does not lead to new insights about the features of gaze behavior in athletes. However, the sequence of fixation locations from the stimulus can be used first to automatically recognize expertise and differences in the scan path and second, given sufficient data, to generate an optimal scan path. Ultimately, this scan path can help one understand important expertise-related fixation locations and their sequence in the gaze signal. Furthermore, with an optimal scan path, one can infer the importance of opponents, teammates (or at least parts of such), or the ball for the decision-making process. By looking at the fixation patches and running an object or person detection, a successful orientation of the scene can be achieved. This

leads to a large amount of data which is advantageous for machine learning as machine learning algorithms show their strengths in regression and the classification of large amounts of data. Even in supervised machine learning algorithms we often face the problem of choosing optimal features because there is no indication as to which set of features can best show the expertise of a class.

Next to supervised learning algorithms, where features need to be selected first, other approaches work in an end-to-end learning fashion where features do not need to be identified beforehand. The most important representatives in this field are the convolutional neural networks (CNNs) and recurrent neural networks (RNN), i.e. bidirectional long short-term memory networks (BiLSTM). CNNs are well used in a range of applications like semantic segmentation and object recognition and can learn to distinguish relevant patterns and shapes or to derive abstract objects. Next to CNNs, RNNs and particularly long short-term memory networks (LSTMs) [192], which can find temporal relationships [193], [194], are also widely used. LSTMs optimize RNNs by minimizing the impact of vanishing and exploding gradients. By using a special function block, LSTMs implement a long short-term memory, which pushes the performance of neural networks. These function blocks allow for the remembering of long-time dependencies and previous information. The network learns which information from the past is important for the current output and which can be forgotten (by a forget gate). As the gaze signal is continuous, LSTMs are predesignated to be used in the analysis of temporal patterns in the gaze signal. Currently, both kinds of machine learning techniques are well used for expertise identification in different domains, e.g. in dentistry education [29], [30] or microsurgery [22], [97]–[99]. Neural networks [30] and supervised learning algorithms [29], [97], [175], [176] have both shown their power in objective expertise identification based on gaze behavior. They found major differences in gaze behavior and could link these differences to different expertise classes. This means both machine learning techniques provide suitable methods to deal with large amounts of data and analysis in a fast, objective, and reproducible way.

In this work, we introduce gazePatchNet which combines the strengths of CNNs to detect latent spatial feature relationships, and BiLSTMs to detect temporal feature relationships in fixation patches. To evaluate gazePatchNet, we conducted a study where we showed participants 360° stimuli of defined soccer game situations from the natural perspective of a goalkeeper on a consumer-grade HTC Vive HMD. The gaze was recorded by the integrated SensoMotoric Instruments (SMI) eye tracker with a frequency of 250 Hz. Each stimulus shows a build-up scene and ends after a pass to the user. We used our model to classify the expertise of our participants into three classes, namely, novice, advanced, and expert. This model is meant to serve as a step in the direction of a perceptual-cognitive training system. If our model is robust enough, the discovered knowledge can be used to identify optimal synthetic scan paths that can then be used to train the gaze behavior of

athletes. The underlying hypothesis is that an improved gaze strategy leads to a more reliable recognition of cues and better decision-making based on these cues.

B.2.2. Methods

Stimulus

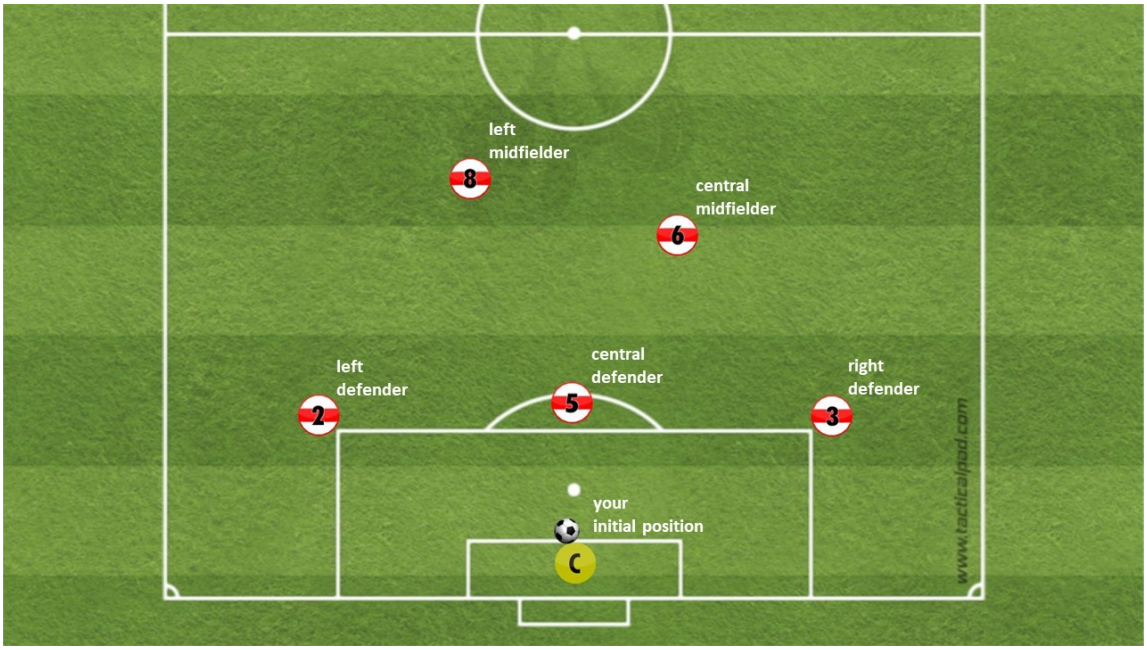

Figure B.2.: Schematic overview of the response options. The sixth option (kick out) is missing.

To show the stimulus video material in virtual reality, we used the SteamVR framework prefab in Unity. SteamVR is an open-source framework that allows common real-time game engines, like Unity, to interface with HMDs. Instead of an artificial recreation of the environment (simulation) within the game engine, we projected realistic footage of 4k omnidirectional videos we captured onto the inside of a sphere around the participant (3840x1920 pixel). This allowed us to display a natural stimulus with high immersion in a realistically mimicked scene. The 360°-footage was captured in cooperation with the German Football Association (DFB) at the training space of the elite youth academy of a German first league club. To capture the footage, we placed a 360° camera at the position of the goalkeeper while 5 teammates and 5 opponents were physically replaying the defined scenarios on the training space. The camera captured the scene with 30 FPS. Each scene was developed based on common scenarios during a match, each

with unique movements. After a video finished, participants had to choose one of six options (five teammates to pass the ball or kick out) to continue the game. In each video, there is one optimal option. This option is counted as one. All five other options are counted as zero. This leads to binary answers in each video. All stimuli were captured on-field and acted out by youth elite players. The plausibility of the scenes, movements, and rating of the decision options were evaluated by an expert trainer team of the DFB. Only stimuli with a single good decision option were included in the experiment. An overview of the options can be seen in Figure B.2 (except "kick out").

Figure B.3.: Example stimulus in equirectangular format.

Data collection

Participants' responses were relayed verbally and finally rated as either right or wrong with only one right decision available per video. The correct decision is a pass to the only teammate who is not covered by an opponent. In total, each participant saw 52 trials, consisting of 26 videos with unique movements, repeated in a different order. Each video trial of each participant counted as one sample.

Participants

Characteristics of all participants can be seen in Table B.1. Data of n=12 experts were collected during a DFB goalkeeper camp, where the DFB gathers the top German elite soccer goalkeepers (U-15 to U-21) for specialized training. These experts are among the top 15 youth elite goalkeepers in Germany. These are the only expert youth players available in Germany. The data of the n=8 advanced

Participants

Class	Age (Mean/SD)	Training hours/week (Mean/SD)	Active years (Mean/SD)
Novice (n=13)	28.64 / 3.72	0.00 / 0.00	1.78 / 5.21
Advanced (n=8)	22.00 / 3.72	4.94 / 0.91	15.50 / 5.77
Expert (n=12)	16.60 / 1.54	8.83 / 4.27	9.16 / 5.04

Table B.1.: Participants summary.

and data of n=13 novice athletes were collected in the university's lab. The advanced players belong to different soccer teams playing in the southern regional league (semi-professional, 4th level) in Germany. The novices have very little to no experience in amateur leagues up to district league with no participation in competitions and no training on a weekly basis.

Procedure

The study was confirmed by the Faculty of Economics and Social Sciences Ethics Committee of the University of Tübingen. After completing a consent form, we started familiarizing the participants with the stimulus presentation and response mode. During the familiarization phase, we showed a sample 360° screenshot of a video on the HMD to allow free exploration of the scene followed by a schematic overview of the field (see Figure B.2). After that, the video scene (see Figure B.3 for an example) was played and stopped (black screen) after the last return pass to the position of the participant. In each scene, we manipulated the color of the ball with a colored dot during the last return pass. This was done in order to force the participants' gaze on the ball during this important phase. As soon as the screen went black, the participant had to report the color of the ball as well as their decision for an option to continue the game. The decision selection is identical to the initial schematic overview of the field (Figure B.2) plus an emergency option to "kick out". This procedure was repeated 5 times.

After this learning phase, we started the first block of 26 trials. The second block contained the same 26 videos, but in a different order. Between the blocks, participants could take off the HMD for a break. Figure B.3 shows a screenshot

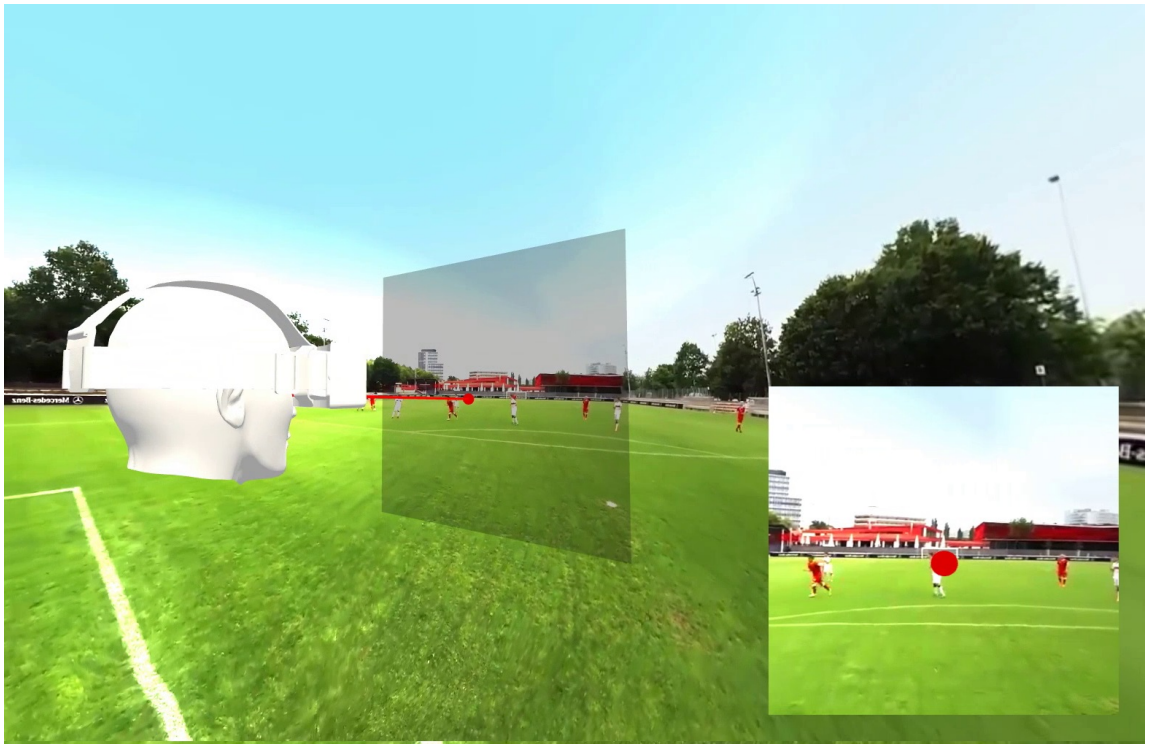

Figure B.4.: Screenshot of an animation showing our system setup. The red line and dot are the gaze signal. The gray rectangle is the field of view of the user inside the HMD. The content of the field of view is shown in the small rectangle on the bottom right side. The users are able to freely move their head/perspective in the scene.

during data collection and Figure B.4 shows a simulation to visualize the setup of the eye-tracking and VR setting.

Image patch extraction

As introduced above, our method (coined gazePatchNet) includes 1) finding latent features in the image patches around the participants' fixations and 2) classifying the scan path as a sequence of the consecutive fixation image patches. The whole process is illustrated in Figure B.5. As not all data collection went smoothly because of slippage of the head-set (too loose) or bad calibration results, we reviewed the gaze signal quality of all samples. We only considered a sample valid if the tracking ratio was higher than 75%. We assigned either class 0 (for novice samples), class 1 (for advanced samples), or class 2 (for expert samples) to each sample. After removing invalid data points, we collected all gaze signal samples for each fixation (timestamp,x,y) and saved them with the corresponding omni-

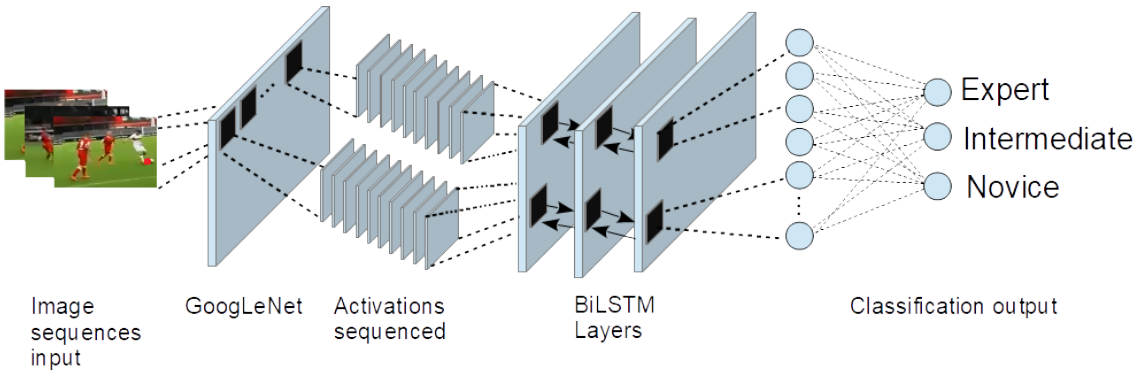

Figure B.5.: gazePatchNet: Our CNN-BiLSTM-based model architecture for expertise classification

directional video file. The fixations were calculated with the vendor's velocity threshold-based event detection filter (I-VT) [37] algorithm using a threshold of 50°/s. We calculated the temporal as well as the spatial center of the fixation based on the averaged gaze signal samples of the fixation. Afterward, we looped over the video file frames to find the corresponding frame by timestamp and cut out an image patch around the fixation on the frame. The size of the patch fits the input size of the input layer of the GoogLeNet CNN (224x224x3 pixels), which we used to extract features later. As soon as we had all the fixation image patches of one trial, we created sequences that fit our BiLSTM. These sequences were essentially fixation image patches in order of their occurrence in the stimulus video.

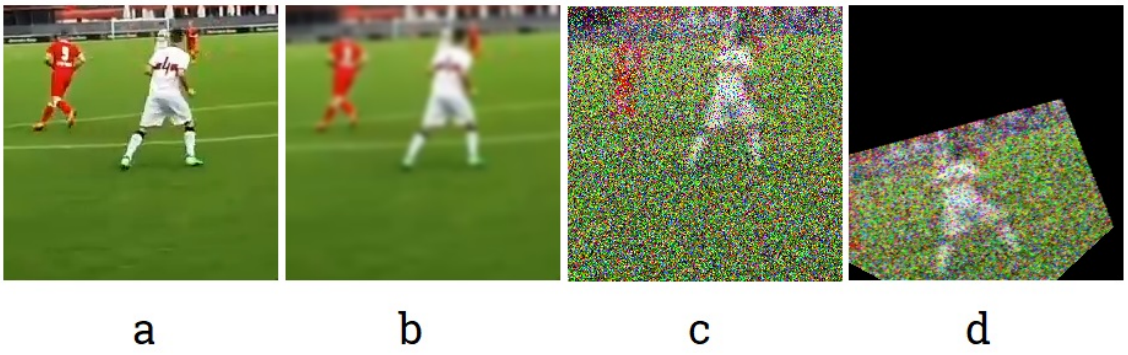

Figure B.6.: Augmentation pipeline. a) shows the original image cut around a fixation, b) shows the image after gaussian blur, c) shows the image after salt & pepper noise addition, and d) shows the transformed image.

Data Augmentation

For each sequence, we computed a new, modified sequence containing the same images. This means we doubled the whole data set by adding the same sequences with the same, just augmented, images. An example is shown in Figure B.6. Figure B.6 a, shows an input image (image cut from the stimulus around a fixation). At first, we applied a random Gaussian blur (Figure B.6, b) and salt & pepper noise (Figure B.6, c). Afterwards, we augmented the images in a randomized manner with geometric transformations [195] (Figure B.6, d). Each image was either rotated by a random factor between -180 and 180 degrees, sheared by a random factor between -15 and 15 degrees, or both, flipped on x- or y-axis or was x- or y-translated between -80 and 80 degrees. These augmentation steps were supposed to make training the model harder in a realistic way. We assumed shear and rotation were real translational variations of the participant's head (whole field of view around fixation). This data augmentation was completed before training in an offline manner. The whole data set was augmented in 135 seconds. LSTMs usually support varying sequence lengths, however, as sequences that are much longer than typical sequences can introduce a lot of padding or discard data because of the padding or truncation of sequences, we removed an average of 20 sequences, about 2% of all sequences. The remaining sequences were sorted by length. This led to more homogeneous padding of the input sequences.

Transfer Learning

To get latent spatial features in the image patches automatically, we used a convolutional neural network (CNN, GoogLeNet) as a feature extractor. The CNN was trained on ImageNet, which has about 1000 classes. Each sequence (the augmented sequences included) was fed to the CNN. We did not use the output layers, as we did not need the classification probabilities for the 1000 classes of ImageNet, but, rather, for three classes of expertise. Instead, we proceeded with transfer learning by grabbing the output of the last activation function (see Figure B.7, the last pooling layer of the GoogLeNet network ($"pool5-7x7_{s1}"$), and added the layers of gazePatchNet (see table B.2). We then adapted the output so that it classified our three expertise groups. By using GoogLeNet as a feature extractor, we simultaneously obtained a feature dimension reduction as our images of 224x224x3 pixels were reduced by the CNN to 1024x1 dimensions. As a result, we achieved not only shape, pattern, and object detection, but also the correct input format for an LSTM by keeping track of the input to the CNN and building sequences of related outputs (activated images).

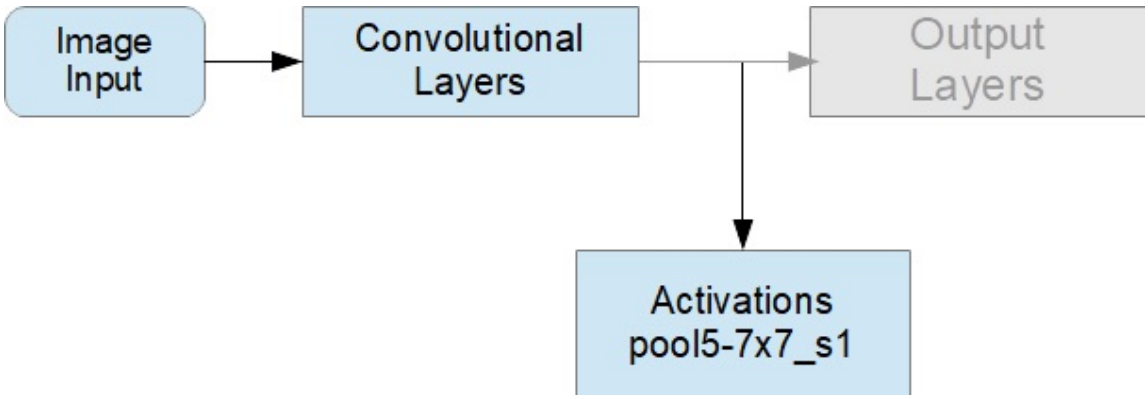

Figure B.7.: Transfer learning for feature selection.

Training and Testing

We trained the model in 33 runs. In each run, the samples of one subject were kept out (leave-one-out validation). The sequences of this subject were used at the end of each run to predict their class. As the model did not see the data before, it was meant to validate the predictive power of the trained model and show how it behaved with totally new data. The data of the remaining 32 subjects were split by a ratio of 70% / 30%. 30% of the data was randomly picked for testing and optimization during training. The remaining 70% of the samples (as well as the augmented samples) were used for training the model.

Model description

Table B.2 shows the structure of the networks' layers. The sequenced activations, containing the selected features, from GoogLeNet were passed to the BiLSTM layers where the temporal relationships were calculated. To input sequences of images into the network, the first layer was a sequence input layer with the same input dimensions (1024) as the output of the activations by the CNN at the last pooling layer (GoogLeNet). As the models with gated recurrent units (GRU) and LSTM layers did not perform well in our tests (between 20-25% lower accuracy), we chose BiLSTM layers as the next part. The BiLSTM layer had 50 hidden units (therefore 4000x1024 input weights, 4000x500 recurrent weights, and 4000x1 biases) and output only the final step. The advantage of BiLSTM layers is that they are fairly generative and take future (forward) and past (backward) states of information into account. After the BiLSTM layer, we added a fully connected layer with 13 hidden units (100x1000 weights and 100x1 biases). To prevent the model from overfitting, we added a dropout layer with a probability of not using a neuron of 50%. As we had three classes to predict, the following fully connected layer had

GazePatchNet

	Name	Type	Activations
1	sequence	Sequence Input	1024
2	bilstm	BiLSTM	1000
3	fc-1	Fully Connected	100
4	dropout	Dropout	100
5	fc-2	Fully Connected	3

Table B.2.: GazePatchNet architecture.

Training Options

Parameter	Value
MiniBatch size	42
Learning rate	4.4e-4
L2-Regularization	8.2e-4
Sequence length	longest sample
Shuffle	no
Validation frequency	52
Validation patience	6
Learning rate schedule	no
Max. epochs	30

Table B.3.: Training Options.

3 hidden units. We took the maximum output to identify the class. To help training converge quickly, we added a softmax layer and calculated the cross-entropy loss for multi-class classification to optimize the model.

Table B.3 provides an overview of the training options. We used grid search to find an optimal hyperparameter set for the whole network [196]. The best set consisted of a mini-batch size of 42, a low learning rate of 4.4e-4, which was not increasing during training time, an L2-regularization of 8.2e-4, to prevent overfitting and a validation frequency that was set to 52 so that the model was validated at every epoch. Validation patience of 6 seemed to be the optimal trade-off between over and underfitting. This means the training was stopped earlier if the loss on the validation set was larger than or equal to the previous smallest loss 6 times in a row. We did not shuffle training and validation data every epoch as we only wanted to use validation data to offer information about the current classification status. The maximum number of epochs for training was set to 100 as longer training results in over or underfitting.

Metrics

We calculated the average/median accuracy over all runs. In each run, 70% of the samples belonged to the training set and 30% to the validation set. We kept one participant out to test how well the model behaved on new, unseen data. As the accuracy is a metric defined by the number of correct predictions divided by the total number of predictions, we could only infer a small amount of information about the model. This was particularly because the samples of the classes used for training and validation were balanced during training, but the distribution of expert, advanced, and novice participants for testing was not. Thus, we also had to consider further performance metrics of the model. The confusion matrix is a sound metric to show the single classes' true and false positives. Similar to the confusion matrix, the following metrics were split into three classes for easy comparison. To gain a deeper performance insight, we showed the receiver operating characteristic (ROC) curve. A ROC curve shows the performance of a classification model at different classification thresholds. Based on the ROC curve, we simply calculated the area under the curve (AUC), which is a common single score and used for comparisons between different models usually on binary classification. Since we split the classes and computed the AUC for each, we compared which classes were predicted most successfully. A score of 1.0 described a perfect skilled model. All scores were calculated by a one-vs-all approach.

B.2.3. Results

The model achieved an average accuracy of 73.11% over 33 runs. For each run, data from one participant was kept out of training and used as test data. We looked

at the data indirectly by describing one trial (one video of a participant) as one sample and classifying these samples as novice, advanced, or expert. This means that some participant samples can be detected as belonging to another group. The distributions of the single samples to different classes can be seen nicely on the confusion matrix in Figure B.8. The accuracy of predicting a novice correctly is at 55.5%. The prediction rate of the advanced class, with an accuracy of 69.4%, is admittedly much higher. And even higher than the advanced class, experts are predicted with an accuracy of 93.4%.

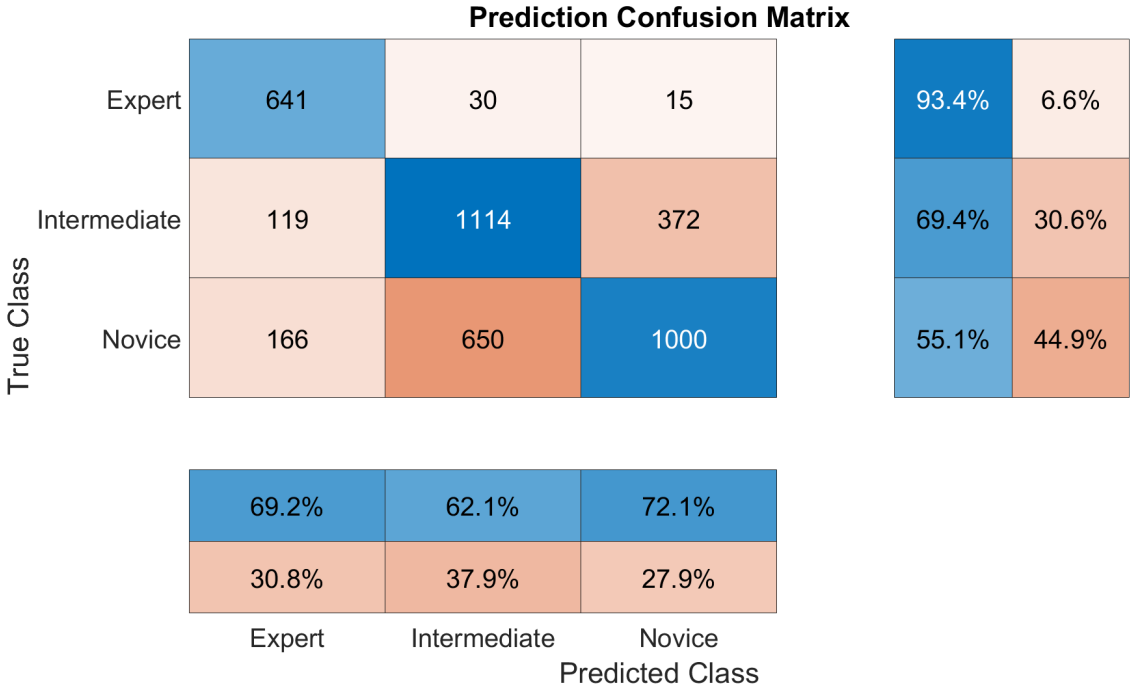

Figure B.8.: Confusion matrix

Out of 1,816 samples of the novice class, 166 samples were predicted as belonging to the expert class and 650 to the advanced class. 1,114 samples were correctly classified as advanced. However, about 1/3 of the advanced samples were predicted falsely, distributed with 372 on novice class and 119 on expert class. 641 of the expert samples were correctly predicted and 30 samples to advanced and 15 to novice class.

Figure B.9 shows three ROC curves with results, corresponding to the confusion matrix. The blue line represents the expert ROC curve. With an AUC of 0.951, the classification is nearly perfect. This corresponds to the confusion matrix values as well. The red line represents the advanced class that does not perform as well as

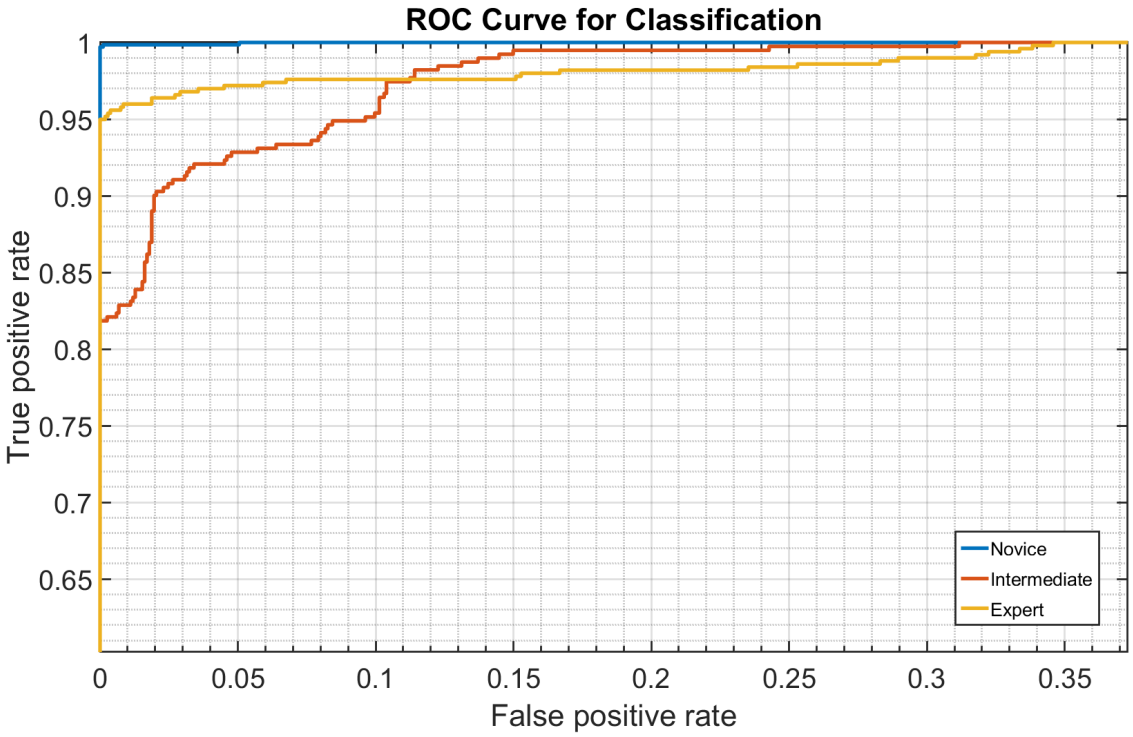

Figure B.9.: ROC-curve for all three classes.

the expert, but still achieves an AUC of 0.833. The yellow line shows the performance of the novice classification, which is a little bit higher than the advanced one with an AUC of 0.871.

B.2.4. Discussion

In this paper, we presented gazePatchNet, a model that, based on a by transfer learning adapted CNN for feature extraction and BiLSTM for temporal dependency identification, automates classification in the broadest sense. We recorded the gaze behavior of soccer goalkeepers during the build-up in a 360° video environment on an HMD and used their fixation image patches on the stimuli as input signals to classify three groups of expertise. The results are promising as we can show, with a relatively small amount of data, that the combination of a CNN, transfer learning, and BiLSTM network is effective in classifying this kind of data. At least the expert and advanced classes are recognizable. However, the novices look more diverse in their behavior and therefore are much harder to predict. The model on average shows great performance, which is reflected by the average accuracy of 73.11% and great sensitivity values visible in the ROC plot. The differences

between experts and the other groups are especially significant.

The classification of the advanced and novice class is about 20% and 40% respectively, lower, but advanced is still doubled when compared to chance-level. This is supposed to increase with more samples for the model to learn from. Our results are well in line with other studies on dynamic tasks, e.g. Bednarik et al. [97] or Eivazi et al [22] who reached a classification accuracy of 66% and 70%, respectively, on medical applications. Both studies predicted the expertise of two skill levels. A more diverse result is found in Castner et al. [29]. Their study predicted the expertise of students from five different semesters alongside experts. With their one-vs-all approach, they mostly reached an accuracy of 37% (chance-level 20%).

In our model, the accuracy of predicting an expert correctly is at 93.4% as this class is the easiest to detect. The prediction rate of the advanced class is much lower with an accuracy of 69.4% because this class is supposed to be the hardest to detect. The accuracy, however, is nearly double the chance level with about 2/3 of the advanced samples classified correctly. Much lower than the advanced, the novices are predicted with an accuracy of 55.1% which is nearly twice as high as the chance level but still 15% lower.

Out of 686 samples of the expert class, only 15 were predicted as belonging to the novice class and 30 to the advanced class. 1,114 samples were correctly classified as advanced, but about 1/3 of the advanced samples were predicted incorrectly. Although interesting, this is no surprise. It may show that the decision boundary for the advanced class does not need to be as robust as many of the advanced participants were gazing like novices and many novices as advanced according to the model. In summary of the performance of the classifications based on the ROC curves, one can state that the samples of all classes were predicted with high certainties and demonstrate the accuracy of a highly skilled predictive model. The average precision value (73.11%), as well as the mean precision of 71.89%, confirm the power of the model.

Looking deeper at the ROC curves, the model performs well in all classes. As samples of advanced players are often predicted as belonging to a novice class and samples of novices players as belonging to an advanced class, it may be necessary to increase the sample size, to robust the decision boundary. In case the model predicts a sample that really belongs to the expert class, it performs this assignment with a high probability of over 93%.

Here, the novice and advanced classes are more difficult to classify. This means that the expert group is a pretty well recognizable group. The advanced and novice groups are more heterogeneous as there are participants that have more/less experience than others. Another reason for this could be that there are missing metrics needed to divide between the two classes properly. This question is typically addressed with the availability of more data. The problem may stem from the small sample size of advanced participants as this group could be too small for the

model to define robust decision boundaries. The fact that experts were barely (15 samples) predicted to be advanced shows that there are clear decision boundaries for advanced and experts. In addition, the cognitive factor is only one of several factors that contribute to expertise. For goalkeepers, for example, it is still most important to be able to block shots on goal. If a goalkeeper can do this extremely well, he may be invited by the DFB even though he could make "worse" decisions after return passes. Conversely, it can also be the case with advanced participants that we have very good decision-makers, but they don't hold as many balls, which is why they are not invited by the DFB. As a result, it is very important to not just test players from different classes but to test players with the assumed highest decision-making skills. For the diagnosis of expertise, we aim to test the best of the very best players and compare them with other expertise groups. We need them to define an optimal behavior. Our expert players are among the 50 most successful young goalkeepers in Germany, which is reflected in the results of our model. A long-range plan is to optimize the training for young players. This work is the first step in that direction. For that, we need to know which behavior is optimal and how we can design training steps for young players to reach this optimal behavior.

The difference in active years/training, and therefore experience, between advanced and expert participants, is much smaller and needs to be finer graded. There may be advanced players with a lot of experience that helps them to perform like experts and there may be experts that don't perform as well because they have much less experience. It is, therefore, not astonishing that some advanced samples are recognized as expert samples. If one assumes that behavior in some samples is better than others, this consequently leads to classifications distributed in different groups. It is more important that the number of classifications of higher-ranked participants into lower classes is minimized in order to depict real expertise.

Instead of providing a description of the behavior of different classes, our model describes a pipeline to find latent features by itself. This circumvents one problem: handcrafted features. The characteristics of handcrafted features may be difficult to teach a user in the form of new behavior based on feature values. Even if the optimal set of features is found, it is difficult to incorporate the findings into a training system. Our model shows a different way of teaching a participant new behavior. As it makes more sense to be able to tell the test person what has to be observed and when and to report it visually, a model should be created that, in the best case, finds an optimal behavior. Based on such information, an optimal behavior for each class can be created and artificially extracted to create information that can be taught to users. A prerequisite will be the analysis of single scan paths, which can be accessed by looking at the fixation image patches.

As the fixation point is currently temporally and spatially averaged, another improvement might be achieved when optimizing the input layer by using an object detection beforehand. Especially when counting in the error rate of the eye tracker and early fixations, some samples might end up directly next to an object and some

directly on it. In this case, the CNN will return different shapes. By using the object as are of interest (AOI) and taking the intersection as input, this behavior can be unified as one can assume that the participant is perceiving the same object in both cases. The CNN can also be optimized. At the moment, this CNN is trained on ImageNet to classify about 1000 classes. By retraining the CNN on a set of 360° videos, with manually labeled teammates, opponents, goals, ball, and free spaces, the intersections of the gaze with AOIs can be advantageous and result in higher classification rates.

Perspectives

As aforementioned, the results already allow for the use of our model as a diagnostic system and as the basis for a training system. The information gathered from this work can be used to model athletes' behavior to personalize new adaptive interfaces that can understand user behaviors based on relevant user information recorded during training. For example, like Wade et al. [197] did for intervention for individuals with autism spectrum disorders. With an objective way to classify the perceptual skill of a person, the first step towards a virtual reality training system (VRTS) with an adaptive design of level difficulty is achieved. With a definition of the perceptual skill of a person and the knowledge of the corresponding skill class, the choice of the difficulty of a level in a VRTS can be adapted automatically. For higher ranked users, the difficulty can be increased by pointing out fewer cues or adapting the stimulus e.g. by placing relevant information outside the foveal area (usage of peripheral vision), designing more crowded scenes (retain overview), or showing highly dynamic situations (faster perception and reaction times). A fundamental work for such a VRTS is to enhance the model with more classes and more participants per class. More data needs to be collected to create a more robust model. A balanced data set would reveal interesting effects on recall and precision and, based on the current performance, might even increase the overall accuracy as the class with the least number of samples has the highest precision values. Different kinds of models also need to be investigated. For feature extraction, a network that is trained on human detection might provide even better results as the head/face and other parts of the human anatomy are potentially considered to be of importance. With 33 participants and an average accuracy of 73.11% on the test set, this model is suitable to be used for this kind of classification.

In a further step, to research the applicability of our model we need to focus on adequate training scenarios. The system can, for example, already be used to create an optimal synthetic scan path. By using the knowledge discovered by our model, one can implement a generative adversarial network. This technique learns to generate new data, in our case a new scan path, with the same statistics as our training set. With enough data to train gazePatchNet to provide strong

robust classes, a synthesized optimal scan path can be created. Should this be possible, it could also become relevant from a practical sports perspective to teach a certain gaze strategy obtained from the generative model. The optimal scan paths identified for each scene could be used to train the gaze behavior of athletes. The underlying hypothesis is that an improved gaze strategy leads to a more reliable recognition of cues and better decision-making based on these cues. To investigate this, however, appropriate training studies are necessary, which must provide information as to whether a) it is feasible for athletes to replace their gaze behavior, developed over years, with a foreign behavior and, if so, whether b) the modification of their gaze behavior also leads to better decision making in the lab. Then the possibility of a corresponding transfer to the field must be checked.

Acknowledgment

This research was supported by the German Football Association (DFB). We thank our colleagues from the DFB who provided insight and expertise that greatly assisted the research. We acknowledge support from the Open Access Publishing Fund of the University of Tübingen. Enkelejda Kasneci is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 - Project number 390727645

C. Gaze-Based Support Timing

This chapter is based on the following publication:

- **B.W. Hosp**, M. S. Yin, P. Haddawy, R. Watcharopas, P. Sa-ngasoongsong, and E. Kasneci (2021). "States of confusion: Eye and Head Tracking Reveal Surgeons' Confusion during Arthroscopic Surgery". In Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada. ACM, New York, NY, USA.

C.1. States of Confusion: Eye and Head Tracking Reveal Surgeons' Confusion During Arthroscopic Surgery

Abstract

During arthroscopic surgeries, surgeons are faced with challenges like cognitive re-projection of the 2D screen output into the 3D operating site or navigation through highly similar tissue. Training of these cognitive processes takes much time and effort for young surgeons but is necessary and crucial for their education. In this study, we want to show how to recognize states of confusion of young surgeons during arthroscopic surgery, by looking at their eye and head movements and feeding them to a machine learning model. With an accuracy of over 94% and detection speed of 0.039 seconds, our model is a step towards online diagnostic and training systems for the perceptual-cognitive processes of surgeons during arthroscopic surgeries.

C.1.1. Introduction

Advancements in computer science have typically been a motor for new applications in fields like medicine. Next to classical imagery techniques like magnetic resonance imaging (MRI) [198] or arthroscopy [199], nowadays, the interaction between surgeons and their patients or instruments are increasingly being investigated. There are a lot of new sources of information, e.g. about the vital parameters of the patient or new perspectives/views of the operating site, which are shown to the physicist. They are all meant to improve the work of the surgeon. However, all these new advancements come with a certain level of complexity. Surgeons need to learn how to operate and benefit from these applications. For example, in arthroscopy, the surgeon needs to transfer the 2D image on the scope output into the 3D tissue of the patient. Information is shown on the screen, but navigation takes place on the operating site with a multidimensional instrument. This translation already poses a challenge.

Even in medical image reading, Brady et al. [200] estimated that the miss rate for interpreting the results correctly, may be up to 30% in some areas of radiology. For arthroscopy, there is no such study, but arthroscopic surgery is a much more complex procedure than image reading, as surgeons are usually under time and success pressure, while working with patients and the stimulus is constantly and dynamically changing. Therefore, ways to teach surgeons to use these new technologies optimally, are as important as the developments of such. This is where human-computer interaction comes into play. Methods of human-computer interaction find their way into the world of medicine. Indeed, there are multiple goals

to pursue. Besides, i.e. touchless interaction techniques [201], the recognition of strategies of surgeons during an operation [151] are investigated. The recognition of skill [110], [153], [202], [203] or states of confusion [204], [205] of surgeons play a central role in interaction design, as they can help to draw a picture of a surgeons' skills and to find weak-spots that need to be focused on in training. This is done to maximize the output of surgeons and to improve their training. Along with confusion, often frustration or disengagement are involved, if the confusion lasts for too long [206]. Pachman et al. [92] summarized different approaches of the last few years and show that multiple ways of detection have been tried, e.g. facial expressions [207], [208] or learners' postures [209]. D'Mello et al. [209] postulated that models based on a single source had high error rates. Thus, later research focused on multiple sources to detect confusion but could not be fully automated, as external judges needed to be involved [210]. Lallé et al. [173] studied various combinations to predict occurrences of confusion. They reported a 61% prediction rate with 193 features. Thus, their system is hardly usable online since the computation of these 193 features takes too much time. Similarly, the model of Shi et al. [211] is hardly usable in an online setting, too, as their model is too complex and thus needs too much computation power and time. Further, they use images on a display, which might cut off environmental influences, thus, preventing the application of natural gaze behavior. Conversely, we use a soft-cadaver in a real surgical setting, which allows the application of natural gaze behavior. We further use a simple but fast and robust model for classification, which allows the usage in an online fashion.

Most often, surgeons need both of their hands for the operation. So new information and interaction techniques need to focus on other modalities than the surgeon's hands. One way to address this is the use of eye-tracking technology. This technique can either be directly used as an interaction method [201] or as an information provider about the skill or current state of the surgeons themselves. And as these devices are getting more ubiquitous, faster, and more accurate, there are ever new possibilities to study the gaze behavior of the subject. Eye tracking can serve as a perceptual-cognitive diagnosis system. The interest in using eye tracking as a research method in medicine is growing rapidly (for an overview see Lévêque et al [212]).

There are even studies that focus on assessment of the impact of training with eye tracking, too [213]–[215]. Wilson et al. [213], i.e found significant differences in completion time when showing young surgeons a video with the gaze signal of an expert during laparoscopy, compared to only showing the plain video of the surgery or allowing a free viewing phase. There are plenty of such studies, showing that the findings of gaze behavior studies can even be used to optimize and/or shorten the training surgeons need to go through. While eye-tracking devices are getting faster and ubiquitous, they produce more data, too. On the one hand, more data means more usable information, but on the other hand, there is a rise

in complexity, too. With more data, there can be more inter-dependencies that are hard to understand and handle, especially with traditional techniques like AOI intersection counts [70], [216]–[219]. To allow the analysis of such big data to be much more complex, there is another very important advancement in computer science that has a heavy impact on medicine. Artificial intelligence is applied in a variety of applications in medicine [220]–[223]. The ever-new potentials of machine learning and especially deep learning enable even more complex tasks to be solved and more data to be analyzed.

In this work, we are focusing on the analysis of data from 15 participants during arthroscopic surgery with so-called soft-cadavers. During arthroscopy, the surgeon is mainly focusing on the output of the arthroscope, which shows a plane 2D view of the arthroscopic camera inside the portal hole of the patient. Surgeons need to rely on these images, while they navigate through tissue and bones. A young surgeon with low experience may get confused during navigation since the structures look pretty similar for untrained surgeons. Expert surgeons can rely on their experience and know which visual clues they can use for navigation. In order to optimize the training of young surgeons, we introduce a real-time ready confusion detection model, that recognizes states of confusion of surgeons during arthroscopic surgeries. With the combination of eye tracking, head tracking, and machine learning methods, we present a highly accurate and fast classification model. Detections of such a model can be used to find weak spots of surgeons in real-time and signal assistive actions to be made.

C.1.2. Methods

Data collection

We collected data of 15 surgeons who are all either members of the Orthopedic Department in the Faculty of Medicine from Mahidol University, Thailand, or in the Orthopedics Surgery Residency Program. All subjects were wearing a TobiiGlasses 2 eye tracker (running at 100 Hz) during arthroscopic surgery of the shoulder on a soft cadaver. The cadaver was placed in front of the surgeon and four feet further away we placed a 4k-screen which shows the output of the scope. During the navigation from the portal hole to the operating side, surgeons were telling verbally where they are and where they go to. They also told when they are confused. This means they can either not tell their current position inside the joint or how to continue for sure. In relation to the beginning of the operation, we measured these points of time, where the surgeon told to be unsure/confused.

Feature space

At first, we synchronized the eye-tracking data with the timing data, by adjusting their timestamps to start at the same time relative to the start of the surgery. This allows us to find the points of time of confusion inside the eye-tracking data. In the next step, we cut out a window around every confusion point (+/- one second before and after the event). These pieces of data are considered as "confusion event" samples and the remaining data with no confusion event as "no event" samples.

Each sample contains the following features:

- point of regard (x, y)
- pupil position (average of both eyes)
- pupil diameter (average of both eyes)
- gyroscope (x, y, z)
- accelerometer (x, y, z)

Classification

To build a random forest model, we split the samples into training and test data sets. This is done in a participant-wise manner, which means, if a subject is picked to belong to the training set, all of their samples belong to the training set. We need to do this, as the model would otherwise learn person-specific, so-called idiosyncratic, features (for further information, see [176]). We followed two different approaches, for testing with unseen data.

The first approach follows a 2/3-strategy. We randomly pick 2/3 of the subjects for training and count the number of confusion event samples for each. Afterward, we collect the same amount of "no event" samples from the same subjects. This means for our training set we have the same amount of confusion event samples as no event samples. This firstly leads to a balanced training set (50% confusion event samples and 50% no event samples) and secondly, to a chance-level of 50%, which allows easy interpretation of the results later.

In the second approach, we want to see whether cross-validation during the training would optimize the results. Thus, we split the training set data by 5-fold cross-validation, which means in every run 1/5 of the data (of the training set) is picked to validate/optimize the model, while 4/5 of the data are used for training the model. After each run, we use the samples of the remaining 1/3 subjects (n=5) to test the classification performance with unseen data. As we want to use our model in an online fashion, we need to test the classification accuracy (with unseen data) and the classification speed as well. We show the online computability by creating a queue, which consists of n=2000 samples. In

our test, we keep reading the gaze signal and add one sample to the queue in each step, while the oldest sample is kicked out of the queue. This means at every state the queue has a total of n=2000 samples. The average of each of the features of all samples inside the queue is now computed. These values are now representing the current content of the queue, which we call delta sample. This delta sample is now given to the trained random forest model and to classify it as a "confusion sample" or "no confusion sample". To infer the average performance time, we measure the computation time of 100 single runs and calculate the average performance time.

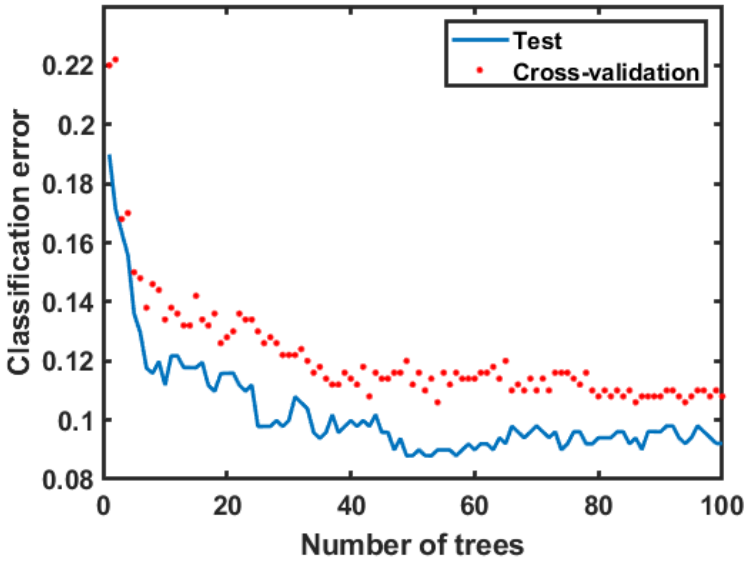

Figure C.1.: Validation with test data vs. validation with cross-validation as function of number of learners.

C.1.3. Results

Out of 1,266,758 samples, we have 7103 samples with a confusion event and 1,259,655 samples with no event. Out of these samples, we collect 7103 confusion samples and 7103 no confusion samples. In every run, we randomly pick 1,000 samples of both to predict their class. The other samples are used for training. We tested our approach - by randomly assigning training and testing data like the aforementioned, 100 times. The average accuracy of the random forest model is 94.2%. According to the accuracy, the average misclassification cost/loss is 0.0595. Figure C.1 shows the development of the loss over all runs as a function of the number of trained trees. The differences are small but noticeable. The approach with the test data set is performing a little bit better than the cross-validation

approach. Test set approach reaches the best performance of the cross-validation approach ($\tilde{0}.11$) already with about 25-30 trees. The optimal loss value for the test approach is reached at 50 trees with a misclassification cost of 0.085.

Figure C.2 shows the confusion matrix which contains the predictions of all 100 runs. In total, we have 50,000 samples for each class. Of class 0 (no event), 47,016 samples out of 50,136 samples were predicted correctly and 3,120 as confusion event samples. Similarly, for class 1 (confusion event), the model predicted 47,023 samples correctly as confusion event and 2,841 samples wrongly as no event. This result is supported, by the average accuracy over all 100 runs of 94.2%.

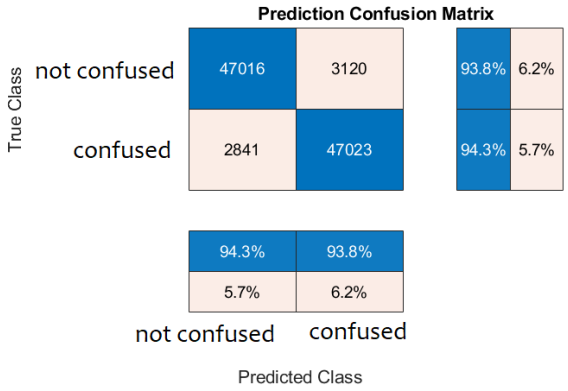

Figure C.2.: Confusion matrix showing number of correctly and falsely predicted samples.

To measure the performance speed of the model, we measured the computing time of each of the 100 runs. On average the prediction takes 0.039 seconds. This corresponds to a frame rate of 25 fps.

C.1.4. Discussion

In this work, we presented a random forest model that is able to classify states of confusion of surgeons during arthroscopic surgery of the shoulder with an accuracy of over 94.2%, by taking only 9 features of the eye and head movement into account. In our calculations, the model was able to provide a prediction of the content of a queue containing n=2000 samples (2 seconds of samples) in 0.039 seconds. This corresponds to the temporal resolution of common head-mounted eye trackers which run at a frame rate between 25-30 fps. The speed may need to be optimized, to allow the application to higher-paced field cameras. But in the scenario of surgery, the speed is not a crucial part, rather, a high detection rate is important. With the detection of confusion states, one can help surgeons to proceed, either pointing out visual clues, which may be used by expert surgeons to

navigate or drawing arrows on the output of the arthroscope which tells the surgeon where to navigate next. Another possible usage of the knowledge of states of confusion can be to augment the whole output by describing the scene by segmenting and labeling each bone or tissue. Or simply name the shown parts in the output. There are multiple ways of supporting the confused surgeon. Depending on the state of expertise, the level of support may be chosen, to allow different skilled surgeons, to train their different weak spots. The different kinds of support can be seen in Figure C.3. a) shows a simple arrow, which tells the surgeon where to go next with the arthroscope. b) shows more support by naming the single party of the output, so the surgeon knows which parts are involved and may remember how to proceed. Figure C.3, c shows a similar output like a), but there are only visual clues highlighted, and d) this help would provide the most support, by segmenting and coloring the single parts in different colors and naming them, accordingly.

Figure C.3.: Different kinds of support for a confused surgeon.

References

[1] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye movements and vision*, Springer, 1967, pp. 171–211.

[2] *Tobii pro - eye tracking for research*, 2015. [Online]. Available: `https://www.tobiipro.com/de/`.

[3] *We are smart eye*. [Online]. Available: `https://smarteye.se/`.

[4] J. Sweller, "Cognitive load theory," in *Psychology of learning and motivation*, vol. 55, Elsevier, 2011, pp. 37–76.

[5] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 symposium on eye-tracking research & applications*, 2010, pp. 141–144.

[6] E. Bozkir, D. Geisler, and E. Kasneci, "Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, IEEE, 2019, pp. 1834–1837.

[7] T. Appel *et al.*, "Cross-participant and cross-task classification of cognitive load based on eye tracking," Ph.D. dissertation, Universität Tübingen, 2021.

[8] T. Appel, C. Scharinger, P. Gerjets, and E. Kasneci, "Cross-subject workload classification using pupil-related measures," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–8.

[9] T. Appel, N. Sevcenko, F. Wortha, *et al.*, "Predicting cognitive load in an emergency simulation based on behavioral and physiological measures," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 154–163.

[10] H. Gao, Z. Lu, V. Demberg, and E. Kasneci, "The index of cognitive activity predicts cognitive processing load in linguistic task," 2021.

[11] I. Chen, C.-C. Chang, *et al.*, "Cognitive load theory: An empirical study of anxiety and task performance in language learning," 2009.

[12] G. R. Bradford, "A relationship study of student satisfaction with learning online and cognitive load: Initial results," *The Internet and Higher Education*, vol. 14, no. 4, pp. 217–226, 2011.

[13] N. J. Castner, "Gaze and visual scanpath features for data-driven expertise recognition in medical image inspection," Ph.D. dissertation, Eberhard Karls Universität Tübingen, 2020.

[14] J. Tanaka and I. Gauthier, "Expertise in object and face recognition," *Psychology of learning and motivation*, vol. 36, pp. 83–125, 1997.

[15] C. P. Murphy, A. Wakefield, P. D. Birch, and J. S. North, "Esport expertise benefits perceptual-cognitive skill in (traditional) sport," *Journal of Expertise*, 2021.

[16] N. Castner, T. Appel, T. Eder, *et al.*, "Pupil diameter differentiates expertise in dental radiography visual search," *PloS one*, vol. 15, no. 5, e0223941, 2020.

[17] T. F. Eder, J. Richter, K. Scheiter, *et al.*, "How to support dental students in reading radiographs: Effects of a gaze-based compare-and-contrast intervention," *Advances in Health Sciences Education*, vol. 26, pp. 159–181, 2021.

[18] D. Geisler, N. Castner, G. Kasneci, and E. Kasneci, "A minhash approach for fast scanpath classification," in *ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–9.

[19] N. Castner, S. Klepper, L. Kopnarski, *et al.*, "Overlooking: The nature of gaze behavior and anomaly detection in expert dentists," in *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, 2018, pp. 1–6.

[20] T. Kübler, S. Eivazi, and E. Kasneci, "Automated visual scanpath analysis reveals the expertise level of micro-neurosurgeons," in *MICCAI workshop on interventional microscopy*, 2015, pp. 1–8.

[21] S. Eivazi, A. Hafez, W. Fuhl, *et al.*, "Optimal eye movement strategies: A comparison of neurosurgeons gaze patterns when using a surgical microscope," *Acta Neurochirurgica*, 2017.

[22] S. Eivazi, M. Slupina, W. Fuhl, H. Afkari, A. Hafez, and E. Kasneci, "Towards automatic skill evaluation in microsurgery," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion*, ACM, 2017, pp. 73–76.

[23] N. Castner, S. Eivazi, K. Scheiter, and E. Kasneci, "Using eye tracking to evaluate and develop innovative teaching strategies for fostering image reading skills of novices in medical training," in *Eye Tracking Enhanced Learning (ETEL2017)*, 2017.

[24] T. C. Kübler, C. Rothe, U. Schiefer, W. Rosenstiel, and E. Kasneci, "Subsmatch 2.0: Scanpath comparison and based on subsequence frequencies," *Behavior research methods*, vol. 49, no. 3, pp. 1048–1064, 2017.

[25] T. C. Kübler, "Algorithms for the comparison of visual scan patterns," Ph.D. dissertation, University of Tübingen, 2017.

[26] T. C. Kübler and E. Kasneci, "Automated comparison of scanpaths in dynamic scenes," in *Pfeiffer, Thies ; Essig, Kai (Hrsg.): Proceedings of the 2nd International Workshop on Solutions for Automatic Gaze Data Analysis 2015 (SAGA 2015)*, 2015.

[27] T. C. Kübler, C. Rothe, U. Schiefer, W. Rosenstiel, and E. Kasneci, "Subsmatch 2.0: Scanpath comparison and based on subsequence frequencies," *Behavior research methods*, vol. 49, no. 3, pp. 1048–1064, 2017.

[28] T. C. Kübler, D. R. Bukenberger, J. Ungewiss, *et al.*, "Towards automated comparison of eye-tracking recordings in dynamic scenes," in *2014 5th European Workshop on Visual Information Processing (EUVIP)*, IEEE, 2014, pp. 1–6.

[29] N. Castner, E. Kasneci, T. Kübler, *et al.*, "Scanpath comparison in medical image reading skills of dental students: Distinguishing stages of expertise development," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–9.

[30] N. Castner, T. C. Kuebler, K. Scheiter, *et al.*, "Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing," in *ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–10.

[31] B. W. Hosp, S. Eivazi, M. Maurer, W. Fuhl, D. Geisler, and E. Kasneci, "Remoteeye: An open-source high-speed remote eye tracker," *Behavior research methods*, pp. 1–15, 2020.

[32] *Powerful eye tracking for pc games*, 2021. [Online]. Available: `https://gaming.tobii.com/`.

[33] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[34] A. Bennett and J. Francis, "The eye as an optical system," *The eye*, vol. 4, pp. 101–131, 1962.

[35] E. D. Guestrin, *Remote, non-contact gaze estimation with minimal subject cooperation*. University of Toronto, 2010.

[36] R. Andersson, M. Nyström, and K. Holmqvist, "Sampling frequency and eye-tracking measures: How speed affects durations, latencies, and more," *Journal of Eye Movement Research*, vol. 3, no. 3, 2010.

[37] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 2000, pp. 71–78.

[38] E. Tafaj, G. Kasneci, W. Rosenstiel, and M. Bogdan, "Bayesian online clustering of eye movement data," in *Proceedings of the symposium on eye tracking research and applications*, 2012, pp. 285–288.

[39] T. Santini, W. Fuhl, T. Kübler, and E. Kasneci, "Bayesian identification of fixations, saccades, and smooth pursuits," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 163–170.

[40] W. Fuhl, N. Castner, and E. Kasneci, "Rule-based learning for eye movement type detection," in *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, 2018, pp. 1–6.

[41] B. W. Hosp, M. S. Yin, P. Haddawy, P. Sa-Ngasoongsong, and E. Kasneci, "Differentiating surgeon expertise solely by eye movement features," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, ACM, 2021.

[42] A. Klostermann and S. Moeinirad, "Fewer fixations of longer duration? expert gaze behavior revisited," *German journal of exercise and sport research*, vol. 50, no. 1, pp. 146–161, 2020.

[43] K. A. Ericsson and J. Smith, *Toward a general theory of expertise: Prospects and limits*. Cambridge University Press, 1991.

[44] C. Bertrand, F. Thullier, *et al.*, "Effects of player position task complexity in visual exploration behavior in soccer.," *International Journal of Sport Psychology*, vol. 40, no. 2, pp. 306–323, 2009.

[45] A. Williams and K. Davids, "Visual search strategy, selective attention, and expertise in soccer," *Research quarterly for exercise and sport*, vol. 69, no. 2, pp. 111–128, 1998.

[46] R. Vaeyens, M. Lenoir, A. M. Williams, L. Mazyn, and R. M. Philippaerts, "The effects of task constraints on visual search behavior and decision-making skill in youth soccer players," *Journal of Sport and Exercise Psychology*, vol. 29, no. 2, pp. 147–169, 2007.

[47] A. M. Williams, K. Davids, L. Burwitz, and J. G. Williams, "Visual search strategies in experienced and inexperienced soccer players," *Research quarterly for exercise and sport*, vol. 65, no. 2, pp. 127–135, 1994.

[48] A. Roca, P. R. Ford, A. P. McRobert, and A. M. Williams, "Identifying the processes underpinning anticipation and decision-making in a dynamic time-constrained task," *Cognitive processing*, vol. 12, no. 3, pp. 301–310, 2011.

[49] A. Roca, P. R. Ford, A. P. McRobert, and A. M. Williams, "Perceptual-cognitive skills and their interaction as a function of task constraints in soccer," *Journal of Sport and Exercise Psychology*, vol. 35, no. 2, pp. 144–155, 2013.

[50] T. Nagano, T. Kato, and T. Fukuda, "Visual behaviors of soccer players while kicking with the inside of the foot," *Perceptual and Motor Skills*, vol. 102, no. 1, pp. 147–156, 2006.

[51] C. Perry, K. Goins, B. Marebwa, T. Singh, and T. Herter, "Improvements in visual search contribute to motor skill learning," *Journal of Sports Science*, vol. 20, pp. 279–287,

[52] G. J. Savelsbergh, J. Van der Kamp, A. M. Williams, and P. Ward, "Anticipation and visual search behaviour in expert soccer goalkeepers," *Ergonomics*, vol. 48, no. 11-14, pp. 1686–1697, 2005.

[53] J. Krzepota, M. Stępiński, and T. Zwierko, "Gaze control in one versus one defensive situations in soccer players with various levels of expertise," *Perceptual and Motor Skills*, vol. 123, no. 3, pp. 769–783, 2016.

[54] A. Williams and K. Davids, "Assessing cue usage in performance contexts: A comparison between eye-movement and concurrent verbal report methods," *Behavior Research Methods, Instruments, & Computers*, vol. 29, no. 3, pp. 364–375, 1997.

[55] G. J. Savelsbergh, M. Onrust, A. Rouwenhorst, and J. Van Der Kamp, "In a four-to-four football tactical position game," *Int. J. Sport Psychol*, vol. 37, pp. 248–264, 2006.

[56] R. Vaeyens, M. Lenoir, A. M. Williams, and R. M. Philippaerts, "Mechanisms underpinning successful decision making in skilled youth soccer players: An analysis of visual search behaviors," *Journal of motor behavior*, vol. 39, no. 5, pp. 395–408, 2007.

[57] R. Cañal-Bruland, S. Lotz, N. Hagemann, J. Schorer, and B. Strauss, "Visual span and change detection in soccer: An expertise study," *Journal of cognitive psychology*, vol. 23, no. 3, pp. 302–310, 2011.

[58] J. S. North, A. M. Williams, N. Hodges, P. Ward, and K. A. Ericsson, "Perceiving patterns in dynamic action sequences: Investigating the processes underpinning stimulus recognition and anticipation skill," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 23, no. 6, pp. 878–894, 2009.

[59] G. J. Savelsbergh, A. M. Williams, J. V. D. Kamp, and P. Ward, "Visual search, anticipation and expertise in soccer goalkeepers," *Journal of sports sciences*, vol. 20, no. 3, pp. 279–287, 2002.

[60] T. Woolley, R. Crowther, K Doma, and J. Connor, "The use of spatial manipulation to examine goalkeepers' anticipation," *Journal of sports sciences*, vol. 33, no. 17, pp. 1766–1774, 2015.

[61] W. F. Helsen and J. L. Starkes, "A multidimensional approach to skilled perception and performance in sport," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 13, no. 1, pp. 1–27, 1999.

[62] T. B. McGuckian, M. H. Cole, and G.-J. Pepping, "A systematic review of the technology-based assessment of visual perception and exploration behaviour in association football," *Journal of Sports Sciences*, vol. 36, no. 8, pp. 861–880, 2018.

[63] R. Kredel, C. Vater, A. Klostermann, and E.-J. Hossner, "Eye-tracking technology and the dynamics of natural gaze behavior in sports: A systematic review of 40 years of research," *Frontiers in psychology*, vol. 8, p. 1845, 2017.

[64] D. Panchuk, S. Vine, and J. N. Vickers, "Eye tracking methods in sport expertise," in *Routledge handbook of sport expertise*, Routledge, 2015, pp. 176–187.

[65] S. Brams, G. Ziv, O. Levin, *et al.*, "The relationship between gaze behavior, expertise, and performance: A systematic review.," *Psychological bulletin*, vol. 145, no. 10, p. 980, 2019.

[66] A. Moran, M. Campbell, and J. Toner, "Exploring the cognitive mechanisms of expertise in sport: Progress and prospects," *Psychology of Sport and Exercise*, vol. 42, pp. 8–15, 2019.

[67] D. T. Mann, A. M. Williams, P. Ward, and C. M. Janelle, "Perceptual-cognitive expertise in sport: A meta-analysis," *Journal of Sport and Exercise Psychology*, vol. 29, no. 4, pp. 457–478, 2007.

[68] D. Harris, M. Wilson, T. Holmes, T. de Burgh, and S. Vine, "Eye movements in sports research and practice: Immersive technologies as optimal environments for the study of gaze behaviour," 2020.

[69] M. Wilson, J. McGrath, S. Vine, J. Brewer, D. Defriend, and R. Masters, "Psychomotor control in a virtual laparoscopic surgery training environment: Gaze control parameters differentiate novices from experts," *Surgical endoscopy*, vol. 24, no. 10, pp. 2458–2464, 2010.

[70] D. Manning, S. Ethell, T. Donovan, and T. Crawford, "How do radiologists do it? the influence of experience and training on searching for chest nodules," *Radiography*, vol. 12, no. 2, pp. 134–142, 2006.

[71] M. R. Wilson, J. S. McGrath, S. J. Vine, J. Brewer, D. Defriend, and R. S. Masters, "Perceptual impairment and psychomotor control in virtual laparoscopic surgery," *Surgical endoscopy*, vol. 25, no. 7, pp. 2268–2274, 2011.

[72] A. Gegenfurtner and M. Seppänen, "Transfer of expertise: An eye tracking and think aloud study using dynamic medical visualizations," *Computers & Education*, vol. 63, pp. 393–403, 2013.

[73] Y. Rong, Z. Akata, and E. Kasneci, "Driver intention anticipation based on in-cabin and driving scene monitoring," in *IEEE Conference on Intelligent Transportation Systems (ITSC), 2020*, 2020.

[74] C. Braunagel, D. Geisler, W. Stolzmann, W. Rosenstiel, and E. Kasneci, "On the necessity of adaptive eye movement classification in conditionally automated driving scenarios," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 19–26.

[75] E. Kasneci, T. C. Kübler, C. Braunagel, W. Fuhl, W. Stolzmann, and W. Rosenstiel, "Exploiting the potential of eye movements analysis in the driving context," in *15. Internationales Stuttgarter Symposium Automobil- und Motorentechnik*, Springer Fachmedien Wiesbaden, 2015.

[76] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel, "Driver-activity recognition in the context of conditionally autonomous driving," in *18th International IEEE Conference on Intelligent Transportation Systems (ITSC 2015)*, 2015.

[77] T. C. Kübler, E. Kasneci, K. Aehling, *et al.*, "Driving with glaucoma: Task performance and gaze movements," *Optometry and Vision Science*, vol. 92, no. 11, pp. 1037–1046, 2015.

[78] E. Kasneci, K. Sippel, K. Aehling, *et al.*, "Driving with binocular visual field loss? a study on a supervised on-road parcours with simultaneous eye and head tracking," *PLoS ONE*, vol. 9, no. 2, e87470, 2014.

[79] T. Kübler, W. Fuhl, E. Wagner, and E. Kasneci, "55 rides: Attention annotated head and gaze data during naturalistic driving," in *Eye-Tracking Research and Applications*, ACM, 2021.

[80] M. Dreißig, M. Baccour, T. Schäck, and E. Kasneci, "Driver drowsiness classification based on eye blink and head movement features using the k-nn algorithm," IEEE, 2020.

[81] E. Bozkir, D. Geisler, and E. Kasneci, "Assessment of driver attention during a safety critical situation in vr to generate vr-based training," in *ACM Symposium on Applied Perception 2019*, 2019. DOI: 10.1145/3343036.3343138.

[82] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci, "Online recognition of driver-activity based on visual scanpath classification," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, 23–36, 2017.

[83] E. Kasneci, "Towards the automated recognition of assistance need for drivers with impaired visual field," Ph.D. dissertation, Universität Tübingen, Wilhelmstr. 32, 72074 Tübingen, 2013.

[84] W. Durward, "Organic psychiatry the psychological consequences of cerebral disorder," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 42, no. 8, p. 772, 1979.

[85] M. Hamilton, "Fish's clinical psychopathology: Signs & symptoms in psychiatry," 1985.

[86] C. Simpson, "Doctors and nurses use of the word confused," *The British Journal of Psychiatry*, vol. 145, no. 4, pp. 441–443, 1984.

[87] C. Wakeley *et al.*, "The faber medical dictionary.," *The Faber Medical Dictionary.*, 1953.

[88] M. Koć-Januchta, T. Höffler, G.-B. Thoma, H. Prechtl, and D. Leutner, "Visualizers versus verbalizers: Effects of cognitive style on learning with texts and pictures–an eye-tracking study," *Computers in Human Behavior*, vol. 68, pp. 170–179, 2017.

[89] D. T. Knoepfle, J. T.-y. Wang, and C. F. Camerer, "Studying learning in games using eye-tracking," *Journal of the European Economic Association*, vol. 7, no. 2-3, pp. 388–398, 2009.

[90] M. Porta, S. Ricotti, and C. J. Perez, "Emotional e-learning through eye tracking," in *Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, 2012, pp. 1–6.

[91] M.-L. Lai, M.-J. Tsai, F.-Y. Yang, *et al.*, "A review of using eye-tracking technology in exploring learning from 2000 to 2012," *Educational research review*, vol. 10, pp. 90–115, 2013.

[92] M. Pachman, A. Arguel, L. Lockyer, G. Kennedy, and J. Lodge, "Eye tracking and early detection of confusion in digital learning environments: Proof of concept," *Australasian Journal of Educational Technology*, vol. 32, no. 6, 2016.

[93] I. Di Leo, K. R. Muis, C. A. Singh, and C. Psaradellis, "Curiosity... confusion? frustration! the role and sequencing of emotions during mathematics problem solving," *Contemporary educational psychology*, vol. 58, pp. 121–137, 2019.

[94] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[95] M. Pal, "Random forest classifier for remote sensing classification," *International journal of remote sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[96] M. Maalouf, "Logistic regression in data analysis: An overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011.

[97] R. Bednarik, S. Eivazi, and H. Vrzakova, "A computational approach for prediction of problem-solving behavior using support vector machines and eye-tracking data," in *Eye Gaze in Intelligent User Interfaces*, Springer, 2013, pp. 111–134.

[98] S. Eivazi and R. Bednarik, "Predicting problem-solving behavior and performance levels from visual attention data," in *Proc. Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI*, 2011, pp. 9–16.

[99] S. Eivazi, R. Bednarik, M. Tukiainen, M. von und zu Fraunberg, V. Leinonen, and J. E. Jääskeläinen, "Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2012, pp. 377–380.

[100] B. W. Hosp, F. Schultz, O. Höner, and E. Kasneci, "Soccer goalkeeper expertise identification based on eye movements," *PloS one*, vol. 16, no. 5, e0251070, 2021.

[101] S. D. Sims and C. Conati, "A neural architecture for detecting user confusion in eye-tracking data," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 15–23.

[102] B. W. Hosp, F. Schultz, E. Kasneci, and O. Höner, "Expertise classification of soccer goalkeepers in highly dynamic decision tasks: A deep learning approach for temporal and spatial feature recognition of fixation image patch sequences," *Frontiers in Sports and Active Living*, vol. 3, p. 183, 2021.

[103] V. Likic, "The needleman-wunsch algorithm for sequence alignment," *Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne*, pp. 1–46, 2008.

[104] S. Nakshathram, R. Duraisamy, and M. Pandurangan, "Sequence-order frequency matrix-sampling and machine learning with smith-waterman (sofm-smsw) for protein remote homology detection," 2021.

[105] J. H. Goldberg and J. I. Helfman, "Visual scanpath representation," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 2010, pp. 203–210.

[106] N. Castner, L. Geßler, D. Geisler, F. Hüttig, and E. Kasneci, "Towards expert gaze modeling and recognition of a user's attention in realtime," *Procedia Computer Science*, vol. 176, 2020.

[107] W. Fuhl, N. Castner, T. Kübler, A. Lotz, W. Rosenstiel, and E. Kasneci, "Ferns for area of interest free scanpath classification," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–5.

[108] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "Visualization of eye tracking data: A taxonomy and survey," in *Computer Graphics Forum*, Wiley Online Library, vol. 36, 2017, pp. 260–284.

[109] T. Zwierko, W. Jedziniak, B. Florkiewicz, *et al.*, "Oculomotor dynamics in skilled soccer players: The effects of sport expertise and strenuous physical effort," *European journal of sport science*, vol. 19, no. 5, pp. 612–620, 2019.

[110] M. S. Yin, P. Haddawy, B. W. Hosp, *et al.*, "A study of expert/novice perception in arthroscopic shoulder surgery," in *Proceedings of the 4th International Conference on Medical and Health Informatics*, 2020, pp. 71–77.

[111] J.-L. Kruger, E. Hefer, and G. Matthew, "Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts," in *Proceedings of the 2013 Conference on Eye Tracking South Africa*, 2013, pp. 62–66.

[112] E. Kocak, J. Ober, N. Berme, and W. S. Melvin, "Eye motion parameters correlate with level of experience in video-assisted surgery: Objective testing of three tasks," *Journal of Laparoendoscopic & Advanced Surgical Techniques*, vol. 15, no. 6, pp. 575–580, 2005.

[113] T. Tien, P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, G.-Z. Yang, and A. Darzi, "Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair," *Surgical endoscopy*, vol. 29, no. 2, pp. 405–413, 2015.

[114] L. Richstone, M. J. Schwartz, C. Seideman, J. Cadeddu, S. Marshall, and L. R. Kavoussi, "Eye metrics as an objective assessment of surgical skill," *Annals of surgery*, vol. 252, no. 1, pp. 177–182, 2010.

[115] N. Merali, D. Veeramootoo, and S. Singh, "Eye-tracking technology in surgical training," *Journal of Investigative Surgery*, 2017.

[116] M. S. Atkins, G. Tien, R. S. Khan, A. Meneghetti, and B. Zheng, "What do surgeons see: Capturing and synchronizing eye gaze for surgery applications," *Surgical innovation*, vol. 20, no. 3, pp. 241–248, 2013.

[117] J. Berry, B. Abernethy, and J. Côté, "The contribution of structured activity and deliberate play to the development of expert perceptual and decision-making skill," *Journal of sport and exercise psychology*, vol. 30, no. 6, pp. 685–708, 2008.

[118] P. Catteeuw, W. Helsen, B. Gilis, and J. Wagemans, "Decision-making skills, role specificity, and deliberate practice in association football refereeing," *Journal of Sports Sciences*, vol. 27, no. 11, pp. 1125–1136, 2009.

[119] B. Abernethy, "Revisiting the relationship between pattern recall and anticipatory skill," *International Journal of Sport Psychology*, vol. 41, pp. 91–106, 2010.

[120] B. Abernethy, J. M. Wood, and S. Parks, "Can the anticipatory skills of experts be learned by novices?" *Research quarterly for exercise and sport*, vol. 70, no. 3, pp. 313–318, 1999.

[121] D. Farrow and B. Abernethy, "Can anticipatory skills be learned through implicit video based perceptual training?" *Journal of sports sciences*, vol. 20, no. 6, pp. 471–485, 2002.

[122] A. M. Williams, P. Ward, J. M. Knowles, and N. J. Smeeton, "Anticipation skill in a real-world task: Measurement, training, and transfer in tennis.," *Journal of Experimental Psychology: Applied*, vol. 8, no. 4, p. 259, 2002.

[123] T. A. DeFanti, G. Dawe, D. J. Sandin, *et al.*, "The starcave, a third-generation cave and virtual reality optiportal," *Future Generation Computer Systems*, vol. 25, no. 2, pp. 169–178, 2009.

[124] M. Wirth, S. Gradl, D. Poimann, *et al.*, "Assessment of perceptual-cognitive abilities among athletes in virtual environments: Exploring interaction concepts for soccer players," in *Proceedings of the 2018 Designing Interactive Systems Conference*, 2018, pp. 1013–1023.

[125] M. Nyström, D. C. Niehorster, R. Andersson, and I. Hooge, "Is the tobii pro spectrum a useful tool for microsaccade researchers?" In *scandinavian workshop on applied eye tracking*, 2018, p. 8.

[126] R. Gray, "Virtual environments and their role in developing perceptual-cognitive skills in sports," *Anticipation and decision making in sport*, pp. 342–358, 2019.

[127] D. J. Harris, J. M. Bird, A. P. Smart, M. R. Wilson, and S. J. Vine, "A framework for the testing and validation of simulated environments in experimentation and training," *Frontiers in Psychology*, vol. 11, p. 605, 2020.

[128] B. Bideau, R. Kulpa, N. Vignais, S. Brault, F. Multon, and C. Craig, "Using virtual reality to analyze sports performance," *IEEE Computer Graphics and Applications*, vol. 30, no. 2, pp. 14–21, 2010.

[129] A. T. Duchowski, V. Shivashankaraiah, T. Rawls, A. K. Gramopadhye, B. J. Melloy, and B. Kanki, "Binocular eye tracking in virtual reality for inspection training," in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 2000, pp. 89–96.

[130] N. S. Uppara, A. A. Mavalankar, and K. Vemuri, "Eye tracking in naturalistic badminton play: Comparing visual gaze pattern strategy in world-rank and amateur player," in *Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*, ACM, 2018, p. 6.

[131] A. I. Grushko and S. V. Leonov, "The usage of eye-tracking technologies in rock-climbing," *Procedia-Social and Behavioral Sciences*, vol. 146, pp. 169–174, 2014.

[132] C. Bard and M. Fleury, "Considering eye movement as a predictor of attainment," *Vision and sport*, vol. 20, pp. 28–41, 1981.

[133] A. T. Bahill, T. LaRitz, *et al.*, "Why can't batters keep their eyes on the ball," *American Scientist*, vol. 72, no. 3, pp. 249–253, 1984.

[134] R. N. Singer, A. M. Williams, S. G. Frehlich, *et al.*, "New frontiers in visual search: An exploratory study in live tennis situations," *Research quarterly for exercise and sport*, vol. 69, no. 3, pp. 290–296, 1998.

[135] J. N. Vickers, "Gaze control in putting," *Perception*, vol. 21, no. 1, pp. 117–132, 1992.

[136] A. Roca, P. R. Ford, and D. Memmert, "Perceptual-cognitive processes underlying creative expert performance in soccer," *Psychological Research*, pp. 1–10, 2020.

[137] M. Dicks, C. Button, and K. Davids, "Examination of gaze behaviors under in situ and video simulation task constraints reveals differences in information pickup for perception and action," *Attention, Perception, & Psychophysics*, vol. 72, no. 3, pp. 706–720, 2010.

[138] K. M. Aksum, L. Magnaguagno, C. T. Bjørndal, and G. Jordet, "What do football players look at? an eye-tracking analysis of the visual fixations of players in 11 v 11 elite football match play," *Frontiers in psychology*, vol. 11, 2020.

[139] A. Roca, P. R. Ford, and D. Memmert, "Creative decision making and visual search behavior in skilled soccer players," *PloS one*, vol. 13, no. 7, e0199381, 2018.

[140] D. Bishop, G. Kuhn, and C. Maton, "Telling people where to look in a soccer-based decision task: A nomothetic approach," 2014.

[141] H. Collewijn, C. J. Erkelens, and R. M. Steinman, "Binocular co-ordination of human horizontal saccadic eye movements.," *The Journal of physiology*, vol. 404, no. 1, pp. 157–182, 1988.

[142] I. Agtzidis, M. Startsev, and M. Dorr, "A ground-truth data set and a classification algorithm for eye movements in 360-degree videos," *arXiv preprint arXiv:1903.06474*, 2019.

[143] L. G. Appelbaum and G. Erickson, "Sports vision training: A review of the state-of-the-art in digital training techniques," *International Review of Sport and Exercise Psychology*, vol. 11, no. 1, pp. 160–189, 2018.

[144] L. Wilkins and L. G. Appelbaum, "An early review of stroboscopic visual training: Insights, challenges and accomplishments to guide future studies," *International Review of Sport and Exercise Psychology*, vol. 13, no. 1, pp. 65–80, 2020.

[145] K. Burris, S. Liu, and L. Appelbaum, "Visual-motor expertise in athletes: Insights from semiparametric modelling of 2317 athletes tested on the nike sparq sensory station," *Journal of Sports Sciences*, vol. 38, no. 3, pp. 320–329, 2020.

[146] D. Klemish, B. Ramger, K. Vittetoe, J. P. Reiter, S. T. Tokdar, and L. G. Appelbaum, "Visual abilities distinguish pitchers from hitters in professional baseball," *Journal of sports sciences*, vol. 36, no. 2, pp. 171–179, 2018.

[147] J. Monson, "Advanced techniques in abdominal surgery.," *British Medical Journal*, vol. 307, no. 6915, pp. 1346–1350, 1993.

[148] F. Hermens, R. Flin, and I. Ahmed, "Eye movements in surgery: A literature review," 2013.

[149] B. Law, M. S. Atkins, A. E. Kirkpatrick, and A. J. Lomax, "Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment," in *Proceedings of the 2004 symposium on Eye tracking research & applications*, 2004, pp. 41–48.

[150] B. Zheng, G. Tien, S. M. Atkins, *et al.*, "Surgeon's vigilance in the operating room," *The American Journal of Surgery*, vol. 201, no. 5, pp. 673–677, 2011.

[151] M. H. Sodergren, F. Orihuela-Espina, J. Clark, A. Darzi, and G.-Z. Yang, "A hidden markov model-based analysis framework using eye-tracking data to characterise re-orientation strategies in minimally invasive surgery," *Cognitive processing*, vol. 11, no. 3, pp. 275–283, 2010.

[152] G. Tien, B. Zheng, and M. S. Atkins, "Quantifying surgeons' vigilance during laparoscopic operations using eyegaze tracking.," in *MMVR*, 2011, pp. 658–662.

[153] N. Ahmidi, G. D. Hager, L. Ishii, G. Fichtinger, G. L. Gallia, and M. Ishii, "Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2010, pp. 295–302.

[154] P. Olsson, *Real-time and offline filters for eye tracking*, 2007.

[155] G. Tien, M. S. Atkins, B. Zheng, and C. Swindells, "Measuring situation awareness of surgeons in laparoscopic training," in *Proceedings of the 2010 symposium on eye-tracking research & applications*, 2010, pp. 149–152.

[156] M. H. Sodergren, F. Orihuela-Espina, J. Clark, J. Teare, G.-Z. Yang, and A. Darzi, "Evaluation of orientation strategies in laparoscopic cholecystectomy," *Annals of surgery*, vol. 252, no. 6, pp. 1027–1036, 2010.

[157] M. S. Atkins, X. Jiang, G. Tien, and B. Zheng, "Saccadic delays on targets while watching videos," in *Proceedings of the symposium on eye tracking research and applications*, 2012, pp. 405–408.

[158] M. H. Sodergren, F. Orihuela-Espina, P. Mountney, *et al.*, "Orientation strategies in natural orifice translumenal endoscopic surgery," *Annals of surgery*, vol. 254, no. 2, pp. 257–266, 2011.

[159] J. Ehlers, C. Strauch, and A. Huckauf, "A view to a click: Pupil size changes as input command in eyes-only human-computer interaction," *International journal of human-computer studies*, vol. 119, pp. 28–34, 2018.

[160] J. McCuaig, M. Pearlstein, and A. Judd, "Detecting learner frustration: Towards mainstream use cases," in *International Conference on Intelligent Tutoring Systems*, Springer, 2010, pp. 21–30.

[161] J. Tauro and R. Pedowitz, "Arthroscopic skills training modalities," in *Motor Skills Training in Orthopedic Sports Medicine*, Springer, 2017, pp. 53–64.

[162] M. Karahan, G. M. Kerkhoffs, P. Randelli, and G. J. Tuijthof, *Effective training of arthroscopic skills*. Springer, 2015.

[163] J. N. Tofte, B. O. Westerlind, K. D. Martin, *et al.*, "Knee, shoulder, and fundamentals of arthroscopic surgery training: Validation of a virtual arthroscopy simulator," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 33, no. 3, pp. 641–646, 2017.

[164] G. J. Tuijthof, M. N. Van Sterkenburg, I. N. Sierevelt, J. Van Oldenrijk, C. N. Van Dijk, and G. M. Kerkhoffs, "First validation of the passport training environment for arthroscopic skills," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 18, no. 2, pp. 218–224, 2010.

[165] R. A. Pedowitz, J. Esch, and S. Snyder, "Evaluation of a virtual reality simulator for arthroscopy skills development," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 18, no. 6, pp. 1–6, 2002.

[166] A. H. Gomoll, G. Pappas, B. Forsythe, and J. J. Warner, "Individual skill progression on a virtual reality simulator for shoulder arthroscopy: A 3-year follow-up study," *The American journal of sports medicine*, vol. 36, no. 6, pp. 1139–1142, 2008.

[167] A. C. Hoyle, C. Whelton, R. Umaar, and L. Funk, "Validation of a global rating scale for shoulder arthroscopy: A pilot study," *Shoulder & Elbow*, vol. 4, no. 1, pp. 16–21, 2012.

[168] N. R. Howells, M. D. Brinsden, R. S. Gill, A. J. Carr, and J. L. Rees, "Motion analysis: A validated method for showing skill levels in arthroscopy," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 24, no. 3, pp. 335–342, 2008.

[169] S. Goyal, M. A. Radi, I. K.-a. Ramadan, and H. G. Said, "Arthroscopic skills assessment and use of box model for training in arthroscopic surgery using sawbones–"fast" workstation," *Sicot-j*, vol. 2, 2016.

[170] S. Erridge, H. Ashraf, S. Purkayastha, A. Darzi, and M. H. Sodergren, "Comparison of gaze behaviour of trainee and experienced surgeons during laparoscopic gastric bypass," *Journal of British Surgery*, vol. 105, no. 3, pp. 287–294, 2018.

[171] K. Rose and R. Pedowitz, "Fundamental arthroscopic skill differentiation with virtual reality simulation," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 31, no. 2, pp. 299–305, 2015.

[172] P. R. DeLucia, D. Preddy, P. Derby, A. Tharanathan, and S. Putrevu, "Eye movement behavior during confusion: Toward a method," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 58, 2014, pp. 1300–1304.

[173] S. Lallé, C. Conati, and G. Carenini, "Predicting confusion in information visualization from eye tracking and interaction data.," in *IJCAI*, 2016, pp. 2529–2535.

[174] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[175] B. W. Hosp, M. S. Yin, P. Haddawy, P. Sa-Ngasoongsong, and E. Kasneci, "Differentiating surgeon expertise solely by eye movement features," *arXiv preprint arXiv:2102.08155v1*, 2021.

[176] B. W. Hosp, F. Schultz, E. Kasneci, and O. Höner, "Eye movement feature classification for soccer expertise identification in virtual reality," *arXiv preprint arXiv:2009.11676*, 2020.

[177] N. Snegireva, W. Derman, J. Patricios, and K. E. Welman, "Eye tracking technology in sports-related concussion: A systematic review and meta-analysis," *Physiological measurement*, vol. 39, no. 12, 12TR01, 2018.

[178] A. Moran, M. Campbell, and D. Ranieri, "Implications of eye tracking technology for applied sport psychology," *Journal of Sport Psychology in Action*, vol. 9, no. 4, pp. 249–259, 2018.

[179] R. M. Discombe and S. T. Cotterill, "Eye tracking in sport: A guide for new and aspiring researchers," *Sport & Exercise Psychology Review*, vol. 11, no. 2, pp. 49–58, 2015.

[180] D. Fegatelli, F. Giancamilli, L. Mallia, A. Chirico, and F. Lucidi, "The use of eye tracking (et) in targeting sports: A review of the studies on quiet eye (qe)," *Intelligent Interactive Multimedia Systems and Services 2016*, pp. 715–730, 2016.

[181] M. T. Chi, P. J. Feltovich, and R. Glaser, "Categorization and representation of physics problems by experts and novices," *Cognitive science*, vol. 5, no. 2, pp. 121–152, 1981.

[182] G. Tenenbaum, T. Sar-El, and M. Bar-Eli, "Anticipation of ball location in low and high-skill performers: A developmental perspective," *Psychology of Sport and Exercise*, vol. 1, no. 2, pp. 117–128, 2000.

[183] T. Romeas, A. Guldner, and J. Faubert, "3d-multiple object tracking training task improves passing decision-making accuracy in soccer players," *Psychology of Sport and Exercise*, vol. 22, pp. 1–9, 2016.

[184] D. Murr, P. Larkin, and O. Höner, "Decision-making skills of high-performance youth soccer players," *German Journal of Exercise and Sport Research*, pp. 1–10, 2020.

[185] D. Marasso, S. Laborde, G. Bardaglio, and M. Raab, "A developmental perspective on decision making in sports," *International Review of Sport and Exercise Psychology*, vol. 7, no. 1, pp. 251–273, 2014.

[186] B. Travassos, D. Araujo, K. Davids, K O'hara, J Leitão, and A Cortinhas, "Expertise effects on decision-making in sport are constrained by requisite response behaviours–a meta-analysis," *Psychology of Sport and Exercise*, vol. 14, no. 2, pp. 211–219, 2013.

[187] G. Appelbaum and G Erickson, "International review of sport and exercise psychology sports vision training: A review of the state-of-the-art in digital training techniques," *Exerc. Psychol*, vol. 11, pp. 160–189, 2016.

[188] J. J. Cummings and J. N. Bailenson, "How immersive is enough? a meta-analysis of the effect of immersive technology on user presence," *Media Psychology*, vol. 19, no. 2, pp. 272–309, 2016.

[189] J. M. Bird, "Ready exerciser one: Examining the efficacy of immersive technologies in the exercise domain," Ph.D. dissertation, Brunel University London, 2019.

[190] M. Slater, "Immersion and the illusion of presence in virtual reality," *British Journal of Psychology*, vol. 109, no. 3, pp. 431–433, 2018.

[191] N. Hagemann, B. Strauss, and R. Cañal-Bruland, "Training perceptual skill by orienting visual attention," *Journal of sport and exercise psychology*, vol. 28, no. 2, pp. 143–158, 2006.

[192] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[193] H. Tian, Y. Tao, S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, vol. 22, no. 3, pp. 1325–1341, 2019.

[194] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3264–3274, 2018.

[195] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[196] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated Machine Learning*, Springer, Cham, 2019, pp. 3–33.

[197] J. Wade, L. Zhang, D. Bian, *et al.*, "A gaze-contingent adaptive virtual reality driving environment for intervention in individuals with autism spectrum disorders," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 1, pp. 1–23, 2016.

[198] K. H. Höhne, H. Fuchs, and S. M. Pizer, *3D imaging in medicine: algorithms, systems, applications*. Springer Science & Business Media, 2012, vol. 60.

[199] R. W. Ike, W. J. Arnold, and K. C. Kalunian, "Arthroscopy in rheumatology: Where have we been? where might we go?" *Rheumatology*, vol. 60, no. 2, pp. 518–528, 2021.

[200] A. P. Brady, "Error and discrepancy in radiology: Inevitable or avoidable?" *Insights into imaging*, vol. 8, no. 1, pp. 171–182, 2017.

[201] A. Mewes, B. Hensen, F. Wacker, and C. Hansen, "Touchless interaction with software in interventional radiology and surgery: A systematic literature review," *International journal of computer assisted radiology and surgery*, vol. 12, no. 2, pp. 291–305, 2017.

[202] S. Speidel, M. Delles, C. Gutt, and R. Dillmann, "Tracking of instruments in minimally invasive surgery for surgical skill analysis," in *International Workshop on Medical Imaging and Virtual Reality*, Springer, 2006, pp. 148–155.

[203] J. Y. Wu, A. Tamhane, P. Kazanzides, and M. Unberath, "Cross-modal self-supervised representation learning for gesture and skill recognition in robotic surgery," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2021.

[204] M. J. Stillman and L. A. Rybicki, "The bedside confusion scale: Development of a portable bedside test for confusion and its application to the palliative medicine population," *Journal of palliative medicine*, vol. 3, no. 4, pp. 449–456, 2000.

[205] Y. Zhou, T. Xu, S. Li, and S. Li, "Confusion state induction and eeg-based detection in learning," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018, pp. 3290–3293.

[206] S. D'Mello and A. Graesser, "Confusion and its dynamics during device comprehension with breakdown scenarios," *Acta psychologica*, vol. 151, pp. 106–116, 2014.

[207] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[208] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29, 2007.

[209] S. D'Mello and A. Graesser, "Automatic detection of learner's affect from gross body language," *Applied Artificial Intelligence*, vol. 23, no. 2, pp. 123–150, 2009.

[210] S. K. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 147–187, 2010.

[211] Q. Shi, L. Warren, G. Saposnik, and J. C. MacDermid, "Confusion assessment method: A systematic review and meta-analysis of diagnostic accuracy," *Neuropsychiatric disease and treatment*, vol. 9, p. 1359, 2013.

[212] L. Lévêque, H. Bosmans, L. Cockmartin, and H. Liu, "State of the art: Eye-tracking studies in medical imaging," *Ieee Access*, vol. 6, pp. 37 023–37 034, 2018.

[213] M. R. Wilson, S. J. Vine, E. Bright, R. S. Masters, D. Defriend, and J. S. McGrath, "Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: A randomized, controlled study," *Surgical endoscopy*, vol. 25, no. 12, pp. 3731–3739, 2011.

[214] S. J. Vine, R. S. Masters, J. S. McGrath, E. Bright, and M. R. Wilson, "Cheating experience: Guiding novices to adopt the gaze strategies of experts expedites the learning of technical laparoscopic skills," *Surgery*, vol. 152, no. 1, pp. 32–40, 2012.

[215] E. A. Krupinski, A. R. Graham, and R. S. Weinstein, "Characterizing the development of visual search expertise in pathology residents viewing whole slide images," *Human pathology*, vol. 44, no. 3, pp. 357–364, 2013.

[216] M. Mackert, S. E. Champlin, K. E. Pasch, and B. D. Weiss, "Understanding health literacy measurement through eye tracking," *Journal of health communication*, vol. 18, no. sup1, pp. 185–196, 2013.

[217] C. Almansa, M. W. Shahid, M. G. Heckman, S. Preissler, and M. B. Wallace, "Association between visual gaze patterns and adenoma detection rate during colonoscopy: A preliminary investigation," *American Journal of Gastroenterology*, vol. 106, no. 6, pp. 1070–1074, 2011.

[218] E. M. Kok, A. B. de Bruin, J. Leppink, J. J. van Merriënboer, and S. G. Robben, "Case comparisons: An efficient way of learning radiology," *Academic radiology*, vol. 22, no. 10, pp. 1226–1235, 2015.

[219] B. S. Kelly, L. A. Rainford, S. P. Darcy, E. C. Kavanagh, and R. J. Toomey, "The development of expertise in radiology: In chest radiograph interpretation,"expert" search pattern may predate "expert" levels of diagnostic accuracy for pneumothorax identification," *Radiology*, vol. 280, no. 1, pp. 252–260, 2016.

[220] P. Szolovits, *Artificial intelligence in medicine*. Routledge, 2019.

[221] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019.

[222] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, S36–S40, 2017.

[223] A. Ramesh, C. Kambhampati, J. R. Monson, and P. Drew, "Artificial intelligence in medicine.," *Annals of the Royal College of Surgeons of England*, vol. 86, no. 5, p. 334, 2004.